

THE BELL JOURNAL OF ECONOMICS AND MANAGEMENT SCIENCE

3



SPRING '73

VOL. 4 NO. 1

cht

1973

EDITOR:

Paul W. MacAvoy, *Massachusetts Institute of Technology*

ASSOCIATE EDITORS:

Dale W. Jorgenson, *Harvard University*

Harvey J. McMains, *American Telephone and Telegraph Company*

William F. Sharpe, *Stanford University*

ASSISTANT TO THE EDITOR:

Leigh E. Tripoli, *American Telephone and Telegraph Company*

Cub - H05437-3 - P5026

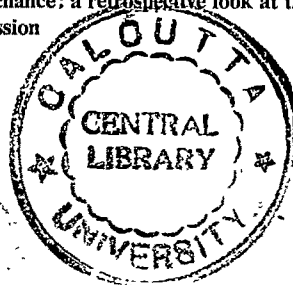
TABLE OF CONTENTS

ARTICLES

- 3 The economics of product safety
W. Y. OI
- 29 An econometric planning model for American Telephone and Telegraph Company
B. E. DAVIS, G. J. CACCAPPOLO,
AND M. A. CHAUDRY
- 57 The behavior of a firm subject to stochastic regulatory review
A. K. KLEVORICK
- 89 The telephone system as a public good: static and dynamic aspects
R. ARTLE AND C. AVEROUS
- 101 A decision analysis with multiple objectives: the Mexico City Airport
R. L. KEENEY
- 118 Pricing decisions of regulated firms: a behavioral approach
P. L. JOSKOW
- 141 The theory of rational option pricing
R. C. MERTON
- 184 Predicting railroad bankruptcies in America
E. I. ALTMAN
- 212 Relative prices on regulated transactions of the natural gas pipelines
P. W. MACAVOY AND R. NOLL
- 235 Market structure, quality and durability
D. LEVHARI AND Y. PELES

SHORT ARTICLES, COMMENTS AND REVIEWS

- 250 The Martin Report and its aftermath
M. MENDELSON
- 270 A dynamic model of the regulated firm with a price adjustment mechanism
E. G. DAVIS
- 283 Resource allocation and the regulated firm: comment
A. B. ATKINSON AND L. WAVERMAN
- 288 Resource allocation and the regulated firm: comment on the comments
E. E. BAILEY
- 293 The profit-performance effects of the separation of ownership from control in large U.S. industrial corporations
J. PALMER
- 304 A simple model of firm behavior under regulation and uncertainty
S. C. MYERS
- 316 Pricing of pollution: the Coase Theorem in the long run
H. E. FRECH III
- 320 Pricing of pollution: reply
R. A. TYBOUT
- 322 Regulated utilities and equipment manufacturers' conspiracies in the electrical power industry
E. D. EMERY
- 338 Depreciation in a simple regulatory model
B. L. JAFFEE
- 343 Television's second chance: a retrospective look at the Sloan Cable Commission
H. J. LEVIN



P-5026

The economics of product safety

Walter Y. Oi

Professor
Graduate School of Management
University of Rochester

The report of the National Commission on Product Safety stated that the American public is being exposed to many unreasonably hazardous products. Further, the post-accident remedies to compensate victims of product-related accidents are inadequate and encourage the continued production of unsafe products. It is believed that a shift in the assignment of liability for accident costs from consumers to sellers would discourage sales of the more hazardous product grades thereby reducing the frequency and severity of home accidents. This conclusion cannot be derived from economic theory. Indeed, under fairly plausible conditions, a shift to producer liability can lead to the increased production of riskier product grades. The theory developed in this paper argues that the demand for a risky product is determined by the full price consisting of the price of the parent risky good plus the imputed expected damage costs. The consumer maximizes utility by demanding that product grade which minimizes his full price. In the concluding section of the paper, attention is directed to the welfare implications of the policy alternatives for the product safety problem.

■ Americans—20 million of them—are injured each year in the home as the result of incidents connected with consumer products. Of the total, 110,000 are permanently disabled, and 30,000 are killed. A significant number could have been spared if more attention had been paid to hazard reduction. The annual cost to the Nation of product-related injuries may exceed \$5.5 billion.¹

This excerpt from the final report of the National Commission on Product Safety (NCPS) describes some of the facts surrounding one of today's leading public policy issues. The public has been reminded of the accident toll resulting from the use of particular consumer products, often by sensational stories about especially gory accidents

1. Introduction

Walter Y. Oi received the Ph.D. degree in economics from the University of Chicago in 1961. He is now a professor in the Graduate School of Management at the University of Rochester. Currently his research is concerned with applied economics in labor and transportation.

This paper was motivated by the Conference on Products Liability (Stanford University, March 1970), sponsored by the American Association of Law Schools and the American Economics Association. Part of the research was conducted while the author was holding the Emma and Carol Roush Distinguished Visiting Scholar position at the Hoover Institution, Stanford University, 1970–1971. Research funds were provided by the Center for Research in Government Policy and Business as well as by the Systems Analysis program at Rochester under U. S. Bureau of Naval Personnel Contract No. RFPN00022-72-R-0139. The views expressed in this paper are those of the author who assumes all responsibility for them and for any errors.

¹ Cf. [20], p. 1. The estimate of 5.5 billion was developed by J. Jepson Wulff. Details of the estimation procedures can be found in the supplementary volumes, [21], pp. 291–362. After reviewing Dr. Wulff's article, I have personally concluded that the estimate of 5.5 billion appears to be high and is clearly subject to a wide margin of error.

inflicting severe physical, psychological, and financial costs. Exploding soda pop bottles, unvented gas heaters, runaway power lawnmowers, unstable high rise bicycles, poisonous lead paints, and shatter-prone sliding glass doors exemplify some of the many *unreasonably* hazardous products identified by the NCPS.² The report stated that the American public is being exposed to many unreasonably (and unnecessarily) hazardous products.

Moreover, the post-accident remedies to compensate accident victims were judged to be inadequate and to encourage manufacturers to persist in the production and sale of unsafe products. The NCPS further concluded that competitive market forces and industry self-regulation are incapable of attaining an acceptable level of safety. In support of these conclusions, they pointed to current laws on legal liability, the high transaction costs of settling damage claims, misleading advertising and imperfect consumer information about product qualities. The Commission believed that efforts to educate consumers on the safe use of potentially dangerous products and to disseminate accurate product information (instructions, warnings, etc.) would do little in reducing the frequency and severity of home accidents. They accordingly recommended the establishment of government regulation over the design and manufacture of products:

Any meaningful and long lasting resolution of problems associated with product safety must start with the product itself. The most basic priority would be to attempt to resolve such problems at the product level before entry into the marketplace, whenever this is possible.³

In the light of their deliberations, the NCPS recommended the establishment of a new Federal agency whose powers would include the regulation of product standards, the compilation and distribution of relevant accident data, and the authority to prohibit the sale of unreasonably hazardous products.⁴

The economic analysis presented in the NCPS report⁵ is quite weak and fails to provide either a logical or an empirical basis for the Commission's recommendations. The Commission appears to be tacitly assuming that the objective of public policy is to minimize the aggregate cost of product-related accidents. But this assumption is

² The Commission endorsed the definition of an "unreasonably hazardous product" proposed by Professor Corwin D. Edwards: "Risks of bodily harm to users are not unreasonable when consumers understand that risks exist, can appraise their probability and severity, know how to cope with them, and voluntarily accept them to get benefits that could not be obtained in less risky ways. When there is a risk of this character, consumers have reasonable opportunity to protect themselves; and public authorities should hesitate to substitute their value judgments about the desirability of the risk for those of the consumers who choose to incur it.

"But preventable risk is not reasonable (a) when consumers do not know that it exists; or (b) when, though aware of it, consumers are unable to estimate its frequency and severity; or (c) when consumers do not know how to cope with it, and hence are likely to incur harm unnecessarily; or (d) when risk is unnecessary in . . . that it could be reduced or eliminated at a cost in money or in the performance of the product that consumers would willingly incur if they knew the facts and were given the choice. Risks that are unreasonable by this definition of unreasonableness seem . . . to be common" [20], p. 11.

³ *Ibid.*, p. 8.

⁴ The full details of the NCPS plan can be found in the appendix to their report in "The Proposed Consumer Product Safety Act" [20].

⁵ *Ibid.*, pp. 67-72.

surely inconsistent with the objective of maximizing the economic welfare of consumers who, of their own free choice, buy soda pop, install sliding glass doors, and operate chain saws.

Several recent studies, notably those by Calabresi and McKean,⁶ provide us with a useful analytic framework for evaluating the welfare implications of different systems of liability assignment for product-related accidents. McKean's analysis begins with Professor R. H. Coase's important theorem,⁷ showing that an optimum allocation of resources which maximizes the total value of production can always be achieved through competitive market forces irrespective of the legal liability assignment if information and transactions were costless. In this ideal world, consumers and producers weigh the costs and benefits before embarking on the consumption and production of more or less hazardous products. Existing institutional arrangements with respect to legal liability (such as the need to establish fault or negligence) can create impediments that prevent parties from engaging in mutually advantageous exchanges which lead to a more optimal allocation of resources. Such impediments clearly exist, and the nonoptimum equilibrium of the economy may involve "too much" or "too little" safety for consumers. The question of "What is a socially optimal level of product safety?" is a difficult one that is not satisfactorily answered either in the NCPS report or in this paper.

My objectives in this paper are more modest. In Section 2, I develop a theory of the demand for a risky product when the product risk is inherent and consumers are fully informed. Under one system of liability assignment, *caveat emptor*, the consumer's choice between different grades of a risky product is shown to depend on the full price including the imputed fully allocated expected cost of damages. When the liability for damages is shifted to producers, sellers are forced to offer a tied package of the risky product plus a full coverage insurance policy. It is shown that a shift from consumer to producer liability can lead to the increased production of the riskier grade. Section 3 of the paper examines how the implications of the formal model are modified when some of the restrictive assumptions are relaxed. Finally, in Section 4, I analyze the welfare implications of the NCPS recommendation for direct government controls over product safety.

■ At the outset, attention is directed to a limiting case in which the product risk is inherent, and all parties are fully informed. The risky product has the characteristic that in any purchase of X units, there is a stable probability distribution, $f(Z: X)$, describing the probability density of obtaining precisely Z good units with $Z \leq X$. The remaining $X - Z$ units are defectives inflicting damage to the consumer. The probability distribution $f(Z: X)$ is assumed to be exogenous and can be described by its mean $\hat{Z}(X)$ and the expected proportion of good

2. A theory of demand for a risky product

⁶ In [6] and [16, 15], respectively. The book by Calabresi [6] offers an excellent legal and economic analysis with emphasis on the costs of auto accidents. McKean's paper [16] served as the focal point of the Conference on Products Liability at Stanford University. He later published a shorter version in the *Quarterly Journal of Economics*, [15].

⁷ In [9].

units π :

$$\int_0^X f(Z: X) dZ = 1, \int_0^X Z f(Z: X) dZ = \hat{Z}(X), \text{ and } \frac{\hat{Z}(X)}{X} = \pi. \quad (1)$$

The consumer is presumed to derive utility from Z good units of the risky product and a riskless product Y , and his utility function can be written,

$$U = U(Z, Y), \text{ with } \frac{dU}{dZ} = U_z \text{ and } \frac{dU}{dY} = U_y. \quad (2)$$

Suppose that each defective unit inflicts a gross damage cost of V dollars consisting of the opportunity price P paid for a defective unit of the parent risky product X and the net damage cost W over and above the price paid for the risky product:

$$V = P + W. \quad (3)$$

The illustrative example of exploding coal due to Buchanan⁸ may help to fix ideas. In this example, firms produce and sell a grade of coal with the property that some unidentifiable lumps may explode when placed in a fire. If the explosions have negligible force, the damage is simply the loss of heat that could have been generated by a good lump; i.e., a bad lump has zero marginal utility. Thus, $V = P$, and the cost is the price paid for a useless lump. If, however, the explosion causes damage to the stove and surroundings, V will exceed P , and the additional net damage cost of $W = V - P$ per defective unit will draw down the resources available to the consumer to demand the other riskless product Y . Finally, in the event that a defective unit has a positive salvage value, the gross damage cost V will be less than P . Thus, the consumer's budget constraint will, in general, be stochastic and depends on the outcome of the random variable Z . Given the identity of equation (3), we can write the budget constraint in two ways:

$$PX + W(X - Z) + Y = M$$

and

$$PZ + V(X - Z) + Y = M, \quad (4)$$

where the price of Y is set equal to one, and M is the consumer's fixed money income per unit time period. Since Z is a random variable, and $Y = M - VX + WZ$, it follows that in a one period analysis, Y will also be random except in the special case of $W = 0$. If $W = 0$, implying $V = P$ —the case of zero net damage cost—the budget constraint is nonstochastic, $Y = M - PX$. One could also treat V as a random variable implying that the severity of an accident is also random. However, in what follows, I shall assume that V and hence W are constant for each consumer but variable across consumers.

□ **The maximization of expected utility.** By letting $Y = M - VX + WZ$, the utility function of equation (2) can be expressed as a function of only one decision variable X , the demand for units of the risky product:

$$U = U(Z, M - VX + WZ). \quad (5)$$

Put in this way, the utility index U will be a random variable, and the expected utility $\hat{U}$ corresponding to any demand for X will be a weighted average determined by the exogenous probability distribution of equation (1):

$$\hat{U} = \int_0^X U(Z, M - VX + WZ)f(Z:X)dZ. \quad (6)$$

A consumer can maximize his expected utility per unit time period by setting the marginal expected utility from an additional unit of the risky product ($d\hat{U}/dX$) equal to zero. Differentiation of equation (6) yields⁹

$$\begin{aligned} \frac{d\hat{U}}{dX} = \int_0^X [U(Z, M - VX + WZ)f_z - VU_z f(Z:X)]dZ \\ + U(X, M - PX)f(X:X). \end{aligned} \quad (7)$$

It is conceptually possible to derive how the optimum demand for units of the risky product X will respond to changes in M , P , and W , but these derivatives are hard to interpret. Moreover, the optimum demand X and the derivatives of the implicit demand function for X are not invariant to all monotonic transformations of the utility function.

□ **Insurable risk and the expected full price.** A consumer who regularly buys the risky product will eventually learn that in any batch of X units, he can expect to get $\pi X = \hat{Z}$ good units and $(1 - \pi)X$ defective units. The cost to the consumer will, on average, be the sum of outlays for the parent risky product, PX , plus the expected net damage cost, $W(X - \hat{Z})$:

$$\hat{C} = PX + W(X - \hat{Z}).$$

By allocating the expected damage cost to good units, we can define the concept of the expected full price P^* as follows:¹⁰

$$P^* = \frac{\hat{C}}{\hat{Z}} = \frac{P}{\pi} + W\left(\frac{1 - \pi}{\pi}\right) \quad (8a)$$

or

$$P^* = \hat{P} + W\gamma, \quad (8b)$$

⁹ Equation (6) is of the form,

$$\hat{U} = \int_0^X \psi(Z, X)dZ,$$

where

$$\psi(Z, X) = U(Z, M - VX + WZ)f(Z:X).$$

The differentiation of an integral equation which can be found in Hillier and Lieberman [12], pp. 606-608, yields

$$\frac{d\hat{U}}{dX} = \int_0^X \left(\frac{d\psi}{dX} \right) dZ + \psi(X, X).$$

In the result shown in equation (7), $f_z = df(Z:X)/dX$ represents the shift in the probability distribution when X is varied, while U_z is the marginal utility of the other riskless product.

¹⁰ In deriving equation (8a), I make use of the identity, $\pi = \hat{Z}/X$. Moreover, it is tacitly assumed that the proportion of good units π is invariant to the scale of purchases.

where $\hat{P} = P/\pi$ will be called the warranty price, and γ is the actuarially fair insurance premium rate for a policy coverage of one dollar:¹¹

$$\gamma = \frac{1 - \pi}{\pi}. \quad (9)$$

These concepts can be illustrated with the aid of a numerical example. Let $P = 1.20$ and $\pi = 0.8$. A consumer who buys 10 units of X at a market cost of $PX = 12$, will, on average, get only 8 units of Z . The warranty price, $\hat{P} = 1.20/0.8 = 1.50$, is thus the expected unit cost of good units when the net damage cost W is equal to zero. If each defective unit inflicts an additional net damage cost of $W = 2.00$, the consumer can insure against the contingent loss by purchasing an insurance policy whose premium cost is $W\gamma = 2.00 \times 0.25 = 0.50$. Thus, the expected full price of a good unit is

$$P^* = \hat{P} + W\gamma = 1.50 + 0.50 = 2.00.$$

Given an inherent and exogenous product risk π , there is clearly a basis for the emergence of an insurance industry that would sell policies against the random occurrence of defective units. In the absence of transaction costs, competition among insurers will, in equilibrium, establish the actuarially fair premium rate γ given by equation (9). The risk averse consumer will demand full coverage protection and pay the fair insurance premium of $W\gamma$ dollars for each unit of Z that he demands.¹²

In the absence of an insurance market, the concept of an expected full price has little meaning when the consumer is assumed to maximize his expected utility in each time period as he does in the model above. If, however, the time horizon is extended, the consumer can still spread the product risk over time and practice self-insurance via a program of sequential purchases combined with borrowing and lending. Returning to the numerical example, our consumer can be imagined to have a desired demand for eight good units per time period. He can always get precisely $Z = 8$ good units if he is pre-

¹¹ A policy paying one dollar in the event of a defective unit that occurs with probability $(1 - \pi)$ entails an expected cost to the insurer of $(1 - \pi)$ dollars. If γ is the premium rate, the insurer's expected revenue is $\gamma\pi$. An actuarially fair premium rate γ is one that yields zero profits:

$$\gamma\pi - (1 - \pi) = 0.$$

The consumer who demands a policy equal to his net damage cost W will incur an insurance premium cost of $W\gamma$ dollars.

¹² The perceptive comments by the anonymous referee for *The Bell Journal* pointed out that I have not adequately dealt with the consumer's attitudes toward risk-bearing. He directs attention to two products having the characteristic that $W\gamma = W'\gamma'$, where $W < W'$ and $\gamma > \gamma'$; i.e., the latter product has a higher proportion of good units, $\pi < \pi'$. In my model, the two products have the same expected net damage cost even though the second product entails a larger financial loss W' . If perfect insurance markets were available to him, the risk averse consumer would demand full coverage protection for both products, and the expected net damage cost is the same for both. If, however, the product risk is uninsurable or if the insurance premium rate includes an administrative "load," I agree with the referee that the latter situation entails a high utilitarian cost.

It should be noticed that $P + V\gamma = \hat{P} + W\gamma$. Hence, if the policy had to cover the opportunity cost of the price paid for the defective unit, full coverage would entail demanding a policy of V dollars. It is thus immaterial whether we defined the full price as $P^* = P + V\gamma$ or $P^* = \hat{P} + W\gamma$.

pared to vary his demand for the parent risky product X . In one period, he might enjoy an especially good run of luck and find that the first 8 units of X he buys are all good units. He realizes, however, that on average, he will have to demand ten units of X at an expected total cost of $\hat{C} = PX + W(X - Z) = 10P + 2W$. His stroke of good fortune means that in this one period, his current outlay of $8P$ is below his expected cost by an amount, $2(P + W)$, and this surplus can be loaned at the market interest rate i . In some later period, he may hit an unusually bad batch that compels him to buy 13 units of X before he gets his 8 good units. His outlay in this later period is $13P + 5W$, and the deficit of $3(P + W)$ could be financed by borrowing at the same interest rate i . With a stable product risk and a sufficiently long time horizon, the borrowing and lending will cancel, and the consumer can always demand precisely 8 good units at a certain cost of $C = ZP^*$.

The presence of a competitive insurance market and/or the ability to spread the product risk through self-insurance enable the consumer to demand any desired quantity of Z at the expected full price P^* . He can thus maximize utility subject to the nonstochastic budget constraint, $P^*Z + Y = M$, yielding the familiar equality between the marginal rate of substitution and relative product prices:

$$-\frac{dY}{dZ} = \frac{U_z}{U_y} = P^*.$$

Holding M constant, the demand for Z will be a function of P^* :

$$Z = D(P^*) \quad \text{and} \quad \eta^* = \frac{P^* D'(P^*)}{Z}, \quad (10)$$

where η^* is the full price elasticity of demand for Z . The demand for the parent risky product X is obtained by recalling that $Z = \pi X$ or $X = Z/\pi$:

$$X = \frac{D(P^*)}{\pi}. \quad (11)$$

Let η_p and η_π denote the elasticities of demand for X with respect to the market price P and the product risk π . It can be shown that¹³

$$\eta_p = \left(\frac{P}{X} \right) \left(\frac{dX}{dP} \right) = \left(\frac{\hat{P}}{P^*} \right) \eta^* \quad (12a)$$

¹³ Differentiation of equation (11) with respect to P yields

$$\frac{dX}{dP} = \frac{1}{\pi} D'(P^*) \left(\frac{dP^*}{dP} \right).$$

From equations (10) and (8a), we have

$$D'(P^*) = \left(\frac{\pi X}{P^*} \right) \eta^*, \quad \frac{dP^*}{dP} = \frac{1}{\pi}.$$

By making these substitutions, we have

$$\frac{dX}{dP} = \left(\frac{X}{\pi P^*} \right) \eta^*.$$

Multiplication by P/X yields equation (12a). Next, if equation (11) is differentiated

and that

$$\eta_\pi = \left(\frac{\pi}{X} \right) \left(\frac{dX}{d\pi} \right) = - \left(1 + \frac{W}{P} \right) \eta_p - 1. \quad (12b)$$

In this model, good units Z are analogous to the ultimate consumption flows developed by Becker.¹⁴

The expected full price P^* will vary across consumers, being higher for those who suffer higher net damage costs W . A one-percent increase in P , the market price of X , will lead to less than a one-percent increase in P^* when $W > 0$. As a consequence, the market price elasticity η_p will be numerically smaller than the full price elasticity η^* . Moreover, consumers who confront larger full prices P^* (because they suffer higher net damage costs W) are likely to exhibit less elastic demands for the parent risky product X .¹⁵

A rise in π , holding P constant, will reduce the full price P^* thereby increasing the demand for good units Z . Such exogenous improvements in product quality (measured here by the product risk π) are ordinarily expected to increase the observable demands for

with respect to π , we get

$$\frac{dX}{d\pi} = \frac{1}{\pi^2} \left[\pi D'(P^*) \left(\frac{dP^*}{d\pi} \right) - D(P^*) \right].$$

Recall that $D(P^*) = Z = \pi X$, and if we multiply by π/X , we get

$$\left(\frac{\pi}{X} \right) \left(\frac{dX}{d\pi} \right) = \left[\frac{D'(P^*)}{X} \right] \left(\frac{dP^*}{d\pi} \right) - 1.$$

From equation (8a), we obtain

$$\frac{dP^*}{d\pi} = - \frac{(P + W)}{\pi^2}.$$

Finally, equation (12a) can be expanded to yield

$$\eta_p = \hat{P} \left[\frac{D'(P^*)}{\pi X} \right] \quad \text{or} \quad \frac{D'(P^*)}{X} = \left(\frac{\pi^2}{P} \right) \eta_p.$$

If we substitute these relationships, we get the result shown in equation (12b).

¹⁴ In [4]. In the Becker model, a household produces an ultimate consumption flow Z by combining purchased market resources X and its own time T . If X units of market resources and T units of time are required to produce one unit of Z , the full price of Z is

$$P_z = P_x X + P_t T,$$

where P_x is the market price of X , and P_t is the implicit cost of time. In my model, the ultimate consumption flow is good units Z that are "produced" by combining purchases of the parent risky product X , $(1/\pi)$ units of X must, on average, be purchased to produce one unit of Z , and damages inflicted by defective units. The insurance premium, $W\gamma$, is the counterpart of Becker's implicit time cost, $P_t T$.

¹⁵ From equations (12a) and (8a), we get

$$A = \frac{\eta_p}{\eta^*} = \frac{P}{P + W(1 - \pi)} < 1.$$

The ratio of the two elasticities approaches zero as W approaches infinity:

$$\left(\frac{W}{A} \right) \left(\frac{dA}{dW} \right) = - (1 - A).$$

If η^* were the same for all consumers, individuals suffering higher net damage costs would exhibit numerically smaller market price elasticities η_p . Becker [4], p. 515, reaches a similar conclusion that products requiring relatively larger time inputs (haircuts, records, golf, etc.) are likely to exhibit smaller elasticities of demand with respect to the prices of the market resource component P_x .

the parent products that are exchanged in markets. However, when one acknowledges the transformation between the parent product X and the ultimate good Z (in our model, $X = Z/\pi$), it is evident that the quantity demanded of X could increase or decrease in response to a rise in π . An improvement in product quality (meaning a larger fraction of good units) will definitely increase the demand for good units Z , but it will raise the demand for units of the parent risky product X if and only if the following inequality is satisfied:

$$-\eta_p > \frac{1}{1 + \frac{W}{P}}.$$

If the inequality were reversed (as it easily could be when the market price elasticity η_p is numerically small), an exogenous improvement in product quality would actually reduce the demand for X .¹⁶

Differences in tastes for Z and Y would be reflected in different full price elasticities η^* . However, even if η^* were the same for all individuals, the observable demand elasticities, η_p and η_π , would still vary across consumers due to the dispersion in net damage costs W .

□ **Consumer liability and choice of product quality.** Although the law on product liability has been changing, the doctrine of *caveat emptor*, wherein the liability for any product-related accident costs is assigned to the consumer, still applies to a wide range of consumer products. The implications of this liability assignment will be analyzed in the context of the model developed in the preceding section.

Suppose that there are two grades of a risky product; a high quality product X_1 has a high fraction π_1 of good units, while a low quality product X_2 has more defective units in each batch. The warranty price of X_1 , $\hat{P}_1 = P_1/\pi_1$, must exceed that of the low quality

¹⁶ A 1-percent increase in π leads to a 1-percent fall in P^* when $W = 0$ and to more than a 1-percent fall in P^* when $W > 0$. Recall that $Z = \pi X$. A 1-percent rise in π must be accompanied by more than a 1-percent increase in the demand for Z before the consumer will expand his demand for the parent product X . In the special case of $W = 0$, the demand elasticity η_π will be positive if and only if the demand for Z and hence X , is relatively elastic, $-\eta^* = -\eta_p > 1$. The elasticity with respect to the product risk, equation (12b) can be rewritten as

$$\eta_\pi = - \left[\frac{1 + \frac{W}{P}}{\frac{W}{1 + \frac{W}{P}(1 - \pi)}} \right] \eta^* - 1.$$

Given the full price elasticity η^* , η_π is an increasing function of the relative net damage cost W/P . Hence, consumers who suffer higher net damage costs W are more likely to increase their demands for X due to an exogenous rise in π .

Professor Gary Becker has suggested to me that this result can be applied to a wider range of quality attributes. It may, for example, explain why the quantity of children demanded falls when the quality of each child is improved. Consider another example in which consumers demand milk for its butter fat content. If a new feed can raise the butter fat content at no additional cost (implying a fall in the full price of butter fat, the ultimate good), the demand for quarts of milk could easily fall as it would if the full price elasticity of demand for butter fat were numerically less than unity.

product, $\hat{P}_2 = P_2/\pi_2$; otherwise, none of X_2 would be demanded.¹⁷ In making his choice between the two grades, the consumer who incurs a positive net damage cost will compare the full prices of good units Z that include both the warranty price $\hat{P}_j$ and the insurance cost $W\gamma_j$. He will thus demand the high quality product if its full price P_1^* is less than the full price P_2^* of a good unit obtained through the purchase of X_2 . Using equation (8b), we see that he will demand X_1 if

$$\hat{P}_1 + W\gamma_1 < \hat{P}_2 + W\gamma_2.$$

By rearranging terms, we get a more convenient expression,

$$W > \frac{\hat{P}_1 - \hat{P}_2}{\gamma_2 - \gamma_1}.$$

There is a critical net damage cost, $W = W^*$, such that $P_1^* = P_2^*$, and the consumer who incurs this critical net damage cost, W^* , will be indifferent between the two products. It is apparent from the above inequality that this critical net damage cost is given by

$$W^* = \frac{\hat{P}_1 - \hat{P}_2}{\gamma_2 - \gamma_1}. \quad (13)$$

In the numerical example of Table 1, the high quality product has a higher warranty price but a lower fair insurance premium. The critical

TABLE 1
A NUMERICAL EXAMPLE FOR TWO GRADES OF A RISKY PRODUCT

ITEM	HIGH QUALITY PRODUCT X_1	LOW QUALITY PRODUCT X_2
1. MARKET PRICE	$P_1 = MC_1 = 1.60$	$P_2 = MC_2 = 1.20$
2. PRODUCT RISK	$\pi_1 = \frac{8}{9}$	$\pi_2 = \frac{8}{10}$
3. WARRANTY PRICE	$\hat{P}_1 = \frac{P_1}{\pi_1} = 1.80$	$\hat{P}_2 = \frac{P_2}{\pi_2} = 1.50$
4. INSURANCE PREMIUM PER DOLLAR OF COVERAGE	$\gamma_1 = \frac{1-\pi_1}{\pi_1} = 0.125$	$\gamma_2 = \frac{1-\pi_2}{\pi_2} = 0.250$
5. CRITICAL NET DAMAGE COST	$W^* = \frac{\hat{P}_1 - \hat{P}_2}{\gamma_2 - \gamma_1} = \frac{1.80 - 1.50}{0.250 - 0.125} = 2.40$	

net damage cost in this example is found to be $W^* = 2.40$. Suppose that individual A suffers a net damage cost of $W = 4.00$ in the event of a defective unit. If A buys the high quality product, he pays a full price $P_1^* = 2.30$, consisting of the warranty price, $\hat{P}_1 = 1.80$, plus the insurance cost, $W\gamma_1 = 4.00 \times 0.125 = 0.50$. Alternatively, if A demands X_2 , he gets a lower warranty price, $\hat{P}_2 = 1.50$, but a larger insurance cost, $W\gamma_2 = 4.00 \times 0.25 = 1.00$, for a full price of $P_2^* = 2.50$. He is thus better off by demanding X_1 because it entails the lower full price. Turn next to individual B who experiences a net

¹⁷ It is assumed that good units Z obtained from either X_1 or X_2 are perfect substitutes. If $\hat{P}_1 < \hat{P}_2$, none of X_2 will be demanded because X_1 offers both a lower adjusted price per good unit and a lower product risk.

damage cost of $W = 2.00$ from each defective unit. By repeating the calculations, we find that the full prices of Z for individual B are $P_1^* = 1.80 + 0.25 = 2.05$, and $P_2^* = 1.50 + 0.50 = 2.00$. Hence, B would prefer to buy X_2 even though its expected damage cost, $W_{\gamma_2} = 0.50$, exceeds that of the high quality product, $W_{\gamma_1} = 0.25$.

The warranty price differential, $(\hat{P}_1 - \hat{P}_2)$ represents the incremental adjusted cost of the high quality product and is determined, in competitive markets by marginal costs and the inherent product risks. Consumers choose between the two grades by weighing the warranty price differential, $(\hat{P}_1 - \hat{P}_2)$, against the differential in insurance costs, $W(\gamma_2 - \gamma_1)$. It should be repeated here that good units Z are assumed to be perfect substitutes, and the net damage costs can be monetized. Those individuals for whom $W > W^*$ will demand X_1 because the benefits from lower expected net damage costs more than offset the higher warranty price for the high quality product. Conversely, those who suffer a net damage cost, $W < W^*$, will realize a lower full price for a good unit Z by demanding the low quality product X_2 .¹⁸ Let $g(W)$ denote the sum of quantities demanded of the risky product X by consumers who incur a net damage cost of precisely W dollars per defective unit.¹⁹ The aggregate market demands for the two grades of the risky product are then given by

$$X_1 = \int_{W^*}^{\infty} g(W)dW \quad \text{and} \quad X_2 = \int_0^{W^*} g(W)dW. \quad (14)$$

A rise in P_1 reduces the demand for the high quality product X_1 in two ways. First, the demand frequencies, $g(W)$, for $W > W^*$, fall as consumers substitute away from Z to the riskless product Y . Second, an increase in P_1 widens the warranty price differential thereby raising the critical net damage cost W^* . Some consumers who previously demanded X_1 will now switch to X_2 because its full price P_2^* will now be lower than the advanced full price P_1^* of the high quality product. Consequently, the demand for X_2 is increased by a rise in P_1 .²⁰ Thus,

¹⁸ The analytic technique of decomposing the full price of an ultimate consumption flow was developed by Becker [4] who emphasized the implicit time cost component. Lancaster [14] employs the same technique, but his components refer to the prices of purchased inputs. R. Muth in a study of the demand for air travel [3] adopts the same technique to divide consumers between different travel modes, specifically jet vs. piston aircraft. Muth assumes that an air traveler chooses the travel mode which provides the lowest full price consisting of the fare F and the implicit cost of travel time VT . The market for jet travel which entailed a higher fare F but a shorter trip time T thus depended in Muth's analysis on the distribution of the implicit hourly cost of travel time V . I have modified this technique to the analysis of demands for a risky product by replacing the implicit time cost by the expected damage cost W_{γ} . Buchanan also argues that under consumer liability, customers will distribute themselves across the several grades of a risky product when he writes: "A person will demand such a low quality product if he places a lower value on risk aversion than the demanders of the high quality product" [5], p. 66. I infer that what he means by "risk aversion" is something akin to my expected damage cost, W_{γ} , and not the usual interpretation of properties of the utility function of wealth where individuals are classified according to their degree of risk aversion or risk preference.

¹⁹ For $W > W^*$, $g(W)$ is the frequency of demands for X_1 by consumers who face a full price of $P_1^* = \hat{P}_1 + W_{\gamma_1}$. Individuals who suffer higher net damage costs W thus confront higher full prices P_1^* . However, for $W < W^*$, $g(W)$ is the frequency of demands for the low quality product X_2 .

²⁰ A rise in P_1 has no effect on the full price, $P_2^* = \hat{P}_2 + W_{\gamma_2}$, for consumers who incur a net damage cost of $W < W^*$. Their demands for X_2 are unaffected;

we have

$$\frac{dX_1}{dP_1} = \int_{W^*}^{\infty} \left[\frac{dg(W)}{dP_1} \right] dW - \frac{g(W^*)}{\pi_1(\gamma_2 - \gamma_1)} < 0 \quad (15a)$$

and

$$\frac{dX_2}{dP_1} = \frac{g(W^*)}{\pi_1(\gamma_2 - \gamma_1)} > 0. \quad (15b)$$

Likewise, a rise in P_2 will reduce the demand for X_2 and increase the demand for X_1 .

Under consumer liability, customers are segregated into two markets according to their net damage cost per defective unit W . The aggregate accident cost α generated by defective units is simply the net damage cost W times the expected number of defective units, $(1 - \pi_1)X_1 = (1 - \pi_1)g(W)$. Since both grades of the risky product will, in general, be sold, α is given by,

$$\alpha = \int_0^{W^*} (1 - \pi_2)Wg(W)dW + \int_{W^*}^{\infty} (1 - \pi_1)Wg(W)dW. \quad (16a)$$

Suppose now that a government agency bans the sale of X_2 because X_2 is, in its opinion, an unreasonably hazardous product. The same result could have been achieved by placing a tax on X_2 , thereby raising its price including the tax to a level P_2' such that $\hat{P}_2' = P_2'/\pi_2 > \hat{P}_1$. Under such a tax, all of the previous demanders of X_2 will now find it advantageous to switch their demands to the high quality coal. The full price of coal to these consumers has thus been raised by the ban, and they are forced to lower indifference curves. Moreover, their demand frequencies $g'(W)$ for the high quality coal X_1 will be less than the pre-ban demand frequencies $g(W)$. If there are constant marginal costs in the production of X_1 , $g(W)$ for $W > W^*$ will be unaffected because P_1^* is unchanged for the previous consumers of high quality coal. The aggregate accident cost α' under the ban is thus given by

$$\alpha' = \int_0^{W^*} (1 - \pi_1)Wg'(W)dW + \int_{W^*}^{\infty} (1 - \pi_1)Wg(W)dW. \quad (16b)$$

A ban will clearly reduce the accident cost. However, the welfare of consumers is *not* determined solely by the size of the accident cost. Given a free choice, the consumers of low quality coal would willingly trade the lower price P_2 for the higher risk of defective units $(1 - \pi_2)$.

□ **Producer liability and resource allocation.** Recent court decisions and the Uniform Commercial Code signal a trend away from consumer liability toward producer liability wherein the manufacturer is held liable for the costs of product-related accidents.²¹ For analytic purposes, let us assume that under producer liability, the seller is

i.e., $dg(W)/dP_1 = 0$, for $W < W^*$. In deriving equations (15a) and (15b), I make use of the result that

$$\frac{dW^*}{dP_1} = \left(\frac{dW^*}{d\hat{P}_1} \right) \left(\frac{d\hat{P}_1}{dP_1} \right) = \left(\frac{1}{\gamma_2 - \gamma_1} \right) \left(\frac{1}{\pi_1} \right).$$

²¹ An excellent review of the legal developments in products liability cases from the viewpoint of an economist is provided by McKean [16]. Another review which places more emphasis on issues of fairness and equity appears in Chapter 7 of the NCPS report [20], pp. 73–80.

legally obligated to compensate customers for all damages that may be inflicted by defective units. Under this regime, the manufacturer is forced into a tie-in sale where he must offer a full coverage insurance policy with each unit of the risky product. His full supply price, $\bar{P}_j$, must now cover both the marginal production cost, $MC_j = P_j$, plus the expected gross damage cost, $V_j\gamma_j = (P_j + W)\gamma_j$. If sellers can discriminate among consumers and charge higher full prices $\bar{P}_j$ to individuals incurring higher net damage costs, the resulting structure of full supply prices $\bar{P}_j$ will be the same as the structure of expected full prices P_j^* that prevailed under consumer liability. But what would happen if sellers were unable to practice "cost-justified" price discrimination and were legally compelled to charge the same full price to all consumers?

The supply price $\bar{P}_j$, quoted by a seller of X_j , must cover both components of the tied package, namely the marginal production cost, $MC_j = P_j$, plus the *mean* expected gross damage cost across all consumers, $\bar{V}_j\gamma_j = (P_j + \bar{W})\gamma_j$. By recalling that $(1 + \gamma_j) = 1/\pi_j$, the full supply price of X_j under producer liability can be written

$$\bar{P}_j = P_j + (P_j + \bar{W})\gamma_j = \hat{P}_j + \bar{W}\gamma_j.$$

Since all damage costs will be paid by sellers, consumers will disregard these costs; they will, accordingly, buy the product grade that offers the lowest full supply price. If both grades are sold, their equilibrium full supply prices (including full coverage insurance policies) must be equal; i.e., $\bar{P}_1 = \bar{P}_2 = \bar{P}$. When all consumers face the same full price of $P^* = \bar{P}$, there is some equilibrium distribution of demand frequencies, $\bar{g}(W)$, representing the amounts of X demanded by consumers who incur a net damage cost of precisely W dollars per defective unit. The *mean* expected net damage cost across all consumers is thus given by

$$\bar{W} = \int_0^\infty W\bar{g}(W)dW. \quad (17)$$

The inability to practice price discrimination, even when the insurance cost component varies across customers, leads to a redistribution of income. Consumers who incur high net damage costs of $W > \bar{W}$ now face a lower full price, $\bar{P} < P^*$, and will thus expand their demands for the risky product. The full price of Z is, however, advanced to those individuals for whom $W < \bar{W}$. They are forced under producer liability to subsidize part of the insurance costs incurred by the former group of high damage cost customers.²²

²² Under producer liability, the full supply price $\bar{P}_j$ is the sum of the marginal production cost, $P_j = MC_j$, plus the *average* expected gross damage cost, $\bar{V}_j\gamma_j = (P_j + \bar{W})\gamma_j$, and *not* the marginal expected gross damage cost. In a competitive market, firms must charge all consumers the same price which covers the expected damage costs. An individual who suffers little from a defective unit, $W < \bar{W}$, is forced via the tie-in sale to "buy" a larger policy $\bar{W}$ than that which he would have demanded under consumer liability. Moreover, if he happens to get a defective unit, he cannot collect $\bar{W}$ dollars, and he thus subsidizes other consumers who incur above average net damage costs. Since the tied package of the parent product plus full coverage insurance is identical for all suppliers of a given product grade, and since firms are unable to discriminate across consumers, the average net damage cost $\bar{W}$ is also the marginal net damage cost to the single supplier.

Some writers have argued that in addition to the aforementioned redistributive effect, producer liability with *no* price discrimination favors the increased production of high quality products thereby reducing accident costs.²³ But a careful analysis of the problem will show that it is not necessarily so.

Suppose that the two products are produced under conditions of constant but different marginal costs. Under consumer liability, competitive market prices are determined by marginal costs, $P_1 = MC_1$ and $P_2 = MC_2$. However, the full prices paid by consumers will vary according to their net damage costs as follows:

$$P_1^* \geq \hat{P}_1 + W^* \gamma_1 \quad (\text{for } W \geq W^*)$$

and

$$P_2^* \leq \hat{P}_2 + W^* \gamma_2 \quad (\text{for } W \leq W^*),$$

where W^* is the critical net damage cost that divides consumers between the two grades.

The imposition of uniform producer liability (meaning no price discrimination) must force one of the two products out of the market because all sellers must charge the same full supply price $\bar{P}$. However, one cannot conclude in advance which product, X_1 or X_2 , will become extinct. If the mean net damage cost across all consumers exceeds the critical net damage costs, $\bar{W} > W^*$, the full supply price by firms producing X_1 will be less than that of firms producing X_2 , and the demand for the low quality product will fall to zero. If, however, $\bar{W} < W^*$, then $\bar{P}_1 > \bar{P}_2$, and consumers who previously demanded the safer product X_1 will now switch their demands to X_2 because they will be fully compensated for all damages inflicted by the more hazardous product.

Suppose, in the example of Table 1, that when all consumers face the same full price of $\bar{P}$, the mean net damage cost is $\bar{W} = 3.60$, which exceeds the critical net damage cost of $W^* = 2.40$. Firms producing the safer product X_1 can supply the tied package of the product plus

It is tacitly assumed in this analysis that all damage costs can be objectively determined, and we do not encounter the moral hazard problem pointed out by Pauly [25].

²³ Quotations from two writers exemplify this argument. Buchanan [5] writes:

"Let us now assume that legal rulings begin to make producing-selling firms fully liable for all damages arising from the usage of either quality of the product. In order to absorb the increased costs of this new liability arrangement, the market price for both qualities of coal will increase, but the market price for the low quality coal will increase much more dramatically. The imposition of strict liability will effectively prohibit firms from marketing a quality of product that inherently involves a relatively larger buyer-user risk. In our example, the low quality product will disappear from the market in an economic sense. It will be impossible for demanders to purchase a quality of product that does not embody full-insurance as a part of the price" [5], p. 66.

McKean reaches a similar conclusion when he writes:

"With the producer facing a higher probability of being liable and with his either carrying liability insurance or paying damages, relatively hazardous designs would be more costly, and the supply curve for such products would decrease. . . . The only thing that would happen to the consumers' position is that he would be denied the opportunity of taking the risk. Since that option would be preferred by some consumers, especially by the poor, this would mean in effect a rise in the price of hazardous products relative to the price of "safe" products, resulting in the end in some shift toward safer products and working to the detriment of the poor" [15], p. 623.

full coverage insurance at a full supply price of $\bar{P}_1 = \bar{P}_1 + \bar{W}\gamma_1 = 1.80 + 0.45 = 2.25$. The sellers of X_2 can offer a lower warranty price but must charge a higher insurance cost because X_2 generates more defective units. Their supply price under producer liability will be $\bar{P}_2 = 1.50 + 0.90 = 2.40$. Hence, the low quality product will be priced out of the market in this case of $\bar{W} > W^*$.

Turn next to the case where, under consumer liability, only a small minority of consumers bought the high quality product. It is probable in this case, that the mean damage cost $\bar{W}$ will fall below the critical net damage cost W^* . If, for example, $\bar{W} = 1.60$, then by repeating the previous calculations, we find the following full supply prices for the two products: $\bar{P}_1 = 1.80 + 0.20 = 2.00$, and $\bar{P}_2 = 1.50 + 0.40 = 1.90$. The tables are now turned, and the high quality product will be priced out of the market, leaving X_2 as the surviving low quality product.

Given constant marginal costs, uniform producer liability will lead to the elimination of one of the two products with the surviving product being determined by the relationship of $\bar{W}$ to W^* . The aggregate accident cost α may rise or fall depending on the elasticities of demand for X and which product is eliminated.²⁴ If there are rising marginal costs, the direction of change will be the same, but the adjustment to equilibrium need not involve the survival of only one product quality. If $\bar{W} > W^*$, the relative demand for high quality coal (X_1/X_2) climbs, increasing the warranty price differential. Conversely, if $W < W^*$, more of the low quality coal will be demanded, narrowing the warranty price differential. In the full equilibrium, both products must sell at the same full price, $\bar{P}_1 = \bar{P}_2$, implying that the

²⁴ The accident cost will definitely increase when the high quality product is eliminated. It is tempting to conclude that if producer liability leads to the disappearance of the riskier product X_2 (meaning $\bar{W} > W^*$), the aggregate accident cost will be reduced. From equation (16), we can disaggregate the accident cost α under consumer liability into three parts according to the consumer's net damage cost:

$$\alpha = a_1 + a_2 + a_3 = \int_0^{W^*} (1 - \pi_2)Wg(W)dW + \int_{W^*}^{\bar{W}} (1 - \pi_1)Wg(W)dW + \int_{\bar{W}}^{\infty} (1 - \pi_1)Wg(W)dW.$$

Under producer liability, only X_1 is sold at the common full price $\bar{P}_1$, the aggregate accident cost $\bar{\alpha}$ is then given by

$$\bar{\alpha} = A_1 + A_2 + A_3 = \int_0^{W^*} (1 - \pi_1)W\bar{g}(W)dW + \int_{W^*}^{\bar{W}} (1 - \pi_1)W\bar{g}(W)dW + \int_{\bar{W}}^{\infty} (1 - \pi_1)W\bar{g}(W)dW.$$

The demand frequencies under producer liability, $\bar{g}(W)$, will differ from those under consumer liability, $g(W)$, because the full price of Z has changed for all three groups as follows: (1) $\bar{P} > P_2^*$ for $0 < W < W^*$, (2) $\bar{P} > P_1^*$ for $W^* < W < \bar{W}$, and (3) $\bar{P} < P_1^*$ for $W > \bar{W}$. Group 3 consumers who incur the highest net damage costs, will enjoy a fall in the full price of Z and will expand their demands for X_1 thereby increasing their contribution to the accident cost. However, the remaining two groups will contract their demands for Z . In fact, group one consumers not only contract their demand for Z but also shift to the safer product. We can thus conclude that $A_1 < a_1$, $A_2 < a_2$, and $A_3 > a_3$, but the magnitude of the differences, $(A_i - a_i)$ cannot be determined without additional knowledge about the price elasticity of demand. Hence, no conclusion can be drawn in advance about the difference between α and $\bar{\alpha}$.

marginal costs must satisfy the following equilibrium condition:²⁵

$$MC_1 = \left(\frac{\pi_1}{\pi_2} \right) MC_2 + \bar{W} \left[\left(\frac{\pi_1}{\pi_2} \right) - 1 \right].$$

Notice that in both cases of constant and rising marginal costs, a shift from consumer to producer liability could easily lead to an increase in the demand for the low quality product X_2 when $\bar{W} < W^*$.²⁶

The conclusion that producer liability will lead to a contraction in the output of more hazardous products appears to rest on an invalid comparison of "market prices." Writers like McKean and Buchanan seem to be comparing the "market prices" P_i for units of the risky product under consumer liability to the full supply prices $\bar{P}_i$ for the product plus full coverage insurance under producer liability. It is surely true that the ratio, $(\bar{P}_i/P_i)$, will be smaller for safer products, suggesting that a shift to producer liability will lead to a smaller relative increase in the "market price" of safer products.²⁷ However, the two "prices" in this ratio pertain to different "goods". A valid comparison should examine the relationship between the full price under consumer liability, P_i^* , and that under producer liability $\bar{P}_i$, where both prices refer to the tied package of the product plus full coverage insurance against damages inflicted by defective units. When one draws the correct comparison by examining the ratio, $(\bar{P}_i/P_i^*)$, it is clear that the demand for the more hazardous product could either rise or fall as a result of shifting to producer liability.

The analysis in this section suggests that the legal assignment of liability can affect the allocation of resources between the two product qualities. The extent to which this is so depends critically on the existence of such impediments as transaction costs or legal enforce-

²⁵ Setting $P_i = MC_i$, the marginal costs must be such that $\bar{P}_1 = \bar{P}_2$.

$$\frac{MC_1}{\pi_1} + \bar{W}\gamma_1 = \frac{MC_2}{\pi_2} + \bar{W}\gamma_2.$$

It can be shown that $(\gamma_2 - \gamma_1) = (\pi_1 - \pi_2)/\pi_1\pi_2$. Using this result and multiplying by π_1 yields the expression shown in the text. Finally, a widening or narrowing of the spread in marginal costs (which, in turn determines the warranty price differential) will affect the level of the full price $\bar{P}$. Moreover, as $\bar{P}$ is varied, the demand frequencies, $\bar{g}(W)$, will also vary leading to an adjustment in the mean net damage cost $\bar{W}$.

²⁶ Most product liability insurance policies place an upper bound, say W_U , on the amount of damages that can be paid to an injured party. Individuals who incur damage costs of $W > W_U$, may continue to demand the high quality product. Limited liability of this type would, however, reduce the mean net damage cost $\bar{W}$ that must be incorporated in the supply price thereby operating to favor the increased production of the more hazardous product.

²⁷ The full supply price under producer liability can be written

$$\bar{P}_i = P_i + (P_i + \bar{W})\gamma_i.$$

We must establish that $(\bar{P}_1/P_1) < (\bar{P}_2/P_2)$ which is simply,

$$1 + \left[1 + \left(\frac{\bar{W}}{P_1} \right) \gamma_1 \right] < 1 + \left[1 + \left(\frac{\bar{W}}{P_2} \right) \gamma_2 \right].$$

Since $\gamma_1 < \gamma_2$ and $P_1 > P_2$, it is apparent that the inequality is satisfied. If consumers have imperfect information about the product risks, these errors will be reflected in their estimates of full prices P_i^* under consumer liability. An overestimate of the safety of the more hazardous product would lower the erroneous estimate for P_2^* ; the relative price of the more hazardous product as Buchanan [5] and McKean [16] suggested.

ment of producer liability with no price discrimination that prevents parties from making mutually advantageous exchanges. Buchanan²⁸ argues that *caveat emptor* is to be preferred to *caveat venditor* (seller beware) because the latter system limits the range of products to safer ones and prevents poorer consumers from maximizing utility by demanding less safe products. Calabresi correctly points out that the two systems of liability assignment are entirely symmetrical if sellers have the same options as consumers. Under *caveat emptor*, consumers can discriminate among sellers and refuse to patronize a seller whose price or product quality is inferior. If the same option were extended to sellers under *caveat venditor*, a manufacturer should be allowed to refuse to sell his product to a consumer who is inferior to his consuming peers because he happens to suffer high damage costs in the event of a defective unit.²⁹ If it is illegal for sellers to practice this kind of discrimination among potential consumers (banks regularly do this when they sell home mortgages), sellers of the risky product are put into the position of offering a legally enforced tie-in sale. In this event, the conclusions set forth in this section hold, and under constant marginal costs, one of the two products will become extinct.

■ The analysis of Section 2 invoked the restrictive assumption that product risks were exogenous, and all parties were fully informed. Actions by consumers and producers can, however, influence both the frequency and severity of product-related accidents. Moreover, the sale of risky products of different qualities may, in part, reflect demands by ill-informed consumers. In this section, I examine how the implications of the analysis of Section 2 are modified when we relax these assumptions.

□ **The profitability of safer products.** In spite of recent trends toward placing strict liability on sellers, consumers must still bear a considerable part of the risks associated with the use of potentially dangerous products. It is alleged that under present legal arrangements, manufacturers lack the incentive to produce reasonably safe products. The NCPS Report points with alarm to cases where firms sell floor furnaces that produce grid temperatures of up to 350 degrees when alternative, safer heating appliances are available, sometimes at lower costs. The implication here is that safer products are simply less profitable for the manufacturer.³⁰

The charge that manufacturers lack incentives to produce safer products is clearly unwarranted when all consumers are fully informed. In Section 2 it was shown that under consumer liability, both grades of the risky product would survive. The analysis can be

3. The production and sale of safer products

²⁸ In [5].

²⁹ This point is amplified in the discussion by Professor Calabresi in the edited transcript of the Products Liability Conference. The reader is referred to [18], pp. 120–122 as well as to Calabresi and Bass [8].

³⁰ A similar point has also been made with respect to product qualities other than safety. In his article on nonprice competition, Heflebower cited examples where products that were “technically identical” often sold at widely varying prices; see [11], p. 183. What appear to be perfect technical substitutes to one observer, may turn out to be very imperfect economic substitutes in the market.

extended to varying degrees of safety.³¹ It is surely reasonable to suppose that the unit price P of the parent risky product will be an increasing function of the proportion of good units π :

$$P = P(\pi) \quad \text{and} \quad \sigma = \left(\frac{\pi}{P} \right) \left(\frac{dP}{d\pi} \right) > 0. \quad (18)$$

The consumer who desires a safer product must be prepared to pay a higher price where σ is the elasticity of the product price with respect to product safety π . Given his net damage cost W , it can be shown that the consumer will maximize utility by demanding a product quality π , with its associated price $P = P(\pi)$, which minimizes his full price P^* of a good unit Z .³² From equation (8a), we have

$$\frac{dP^*}{d\pi} = \left(\frac{P}{\pi^2} \right) \left[\sigma - \left(1 + \frac{W}{P} \right) \right]. \quad (19)$$

An improvement in product quality, meaning a rise in π , affects the full price P^* in two opposing ways. First, P^* will rise because the consumer must pay a higher price for the safer product; the magnitude of this effect is given by σ , the price elasticity of product quality. Second, a rise in π operates to reduce P^* by lowering the expected net damage cost, and the magnitude of this effect is de-

³¹ Spengler [26] introduced the concept of an index of safety S defined over the interval zero to one with higher values indicating more safety. His index has a natural counterpart in simple examples where safety is entirely described by one parameter such as the proportion of good units π . If, however, one must consider several dimensions such as the frequency and severity of accidents, it becomes necessary to transform points from a multi-dimensional space to the zero to one interval. One obvious procedure is to attach monetary values to all points in the complex space describing accidents, but it may be difficult to arrive at these values.

³² Recall that $Z = \pi X$, and the full price can be written

$$P^* = \hat{P} + W_Y = \left(\frac{P + W}{\pi} \right) - W.$$

We then form the Lagrangian expression,

$$L = U(\pi X, Y) + \lambda[M - P^* \pi X - Y].$$

The first-order conditions for a constrained maximum are given by

$$\frac{dL}{dX} = \pi U_x - \lambda \pi P^* = 0,$$

$$\frac{dL}{dY} = U_y - \lambda = 0,$$

and

$$\frac{dL}{d\pi} = X U_x - \lambda X P^* - \lambda \pi X \left(\frac{dP^*}{d\pi} \right) = 0.$$

The first two equations determine the optimum demands for X and Y . Moreover, if the first is substituted into the third equation, we see that $(dL/d\pi)$ will be equal to zero when $(dP^*/d\pi) = 0$; i.e., utility is maximized choosing π so as to minimize the full price P^* of a good unit. Differentiation of P^* with respect to π yields

$$\begin{aligned} \frac{dP^*}{d\pi} &= \left(\frac{1}{\pi^2} \right) \left[\pi \left(\frac{dP}{d\pi} \right) - (P + W) \right] \\ &= \left(\frac{P}{\pi^2} \right) \left[\left(\frac{\pi}{P} \right) \left(\frac{dP}{d\pi} \right) - \left(\frac{P + W}{P} \right) \right]. \end{aligned}$$

This last result is used to arrive at equation (19).

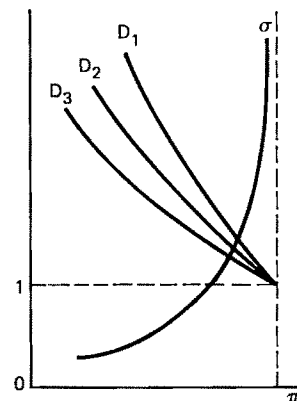
scribed by the relative damage cost, $\left(1 + \frac{W}{P}\right)$. The optimal product quality π for a particular consumer is obtained when the price elasticity of buying a safer product, σ , is just equal to his relative damage cost. Hence, in equilibrium, we have

$$\sigma = 1 + \frac{W}{P}. \quad (20)$$

In Figure 1, σ and $\left(1 + \frac{W}{P}\right)$ are plotted as functions of product risk π . If the marginal cost of superior quality climbs at an increasing rate, the price elasticity of quality, σ , will be an increasing function of the proportion of good units π .³³ Given that $W > 0$, the relative damage cost falls because P is an increasing function of π . The position of the relative damage cost curve depends, however, on the individual's particular net damage cost W per defective unit. In Figure 1, the curve D_1 corresponds to the highest damage cost, and D_3 to the lowest. The optimal product quality is thus determined by the intersection of the D and σ curves. Setting aside indivisibilities, developmental costs, and marketing expenses, this model implies a wide range of product qualities bounded only by the dispersion of net damage costs.³⁴

The potential profitability of safer products under producer liability will depend on the distribution of net damage costs W across consumers. In the case of two product grades, the analysis of Section 2 showed that when $\bar{W} < W^*$, uniform producer liability favored the increased production of the riskier product. The range of product

FIGURE 1
PRODUCT QUALITY



³³ I have tacitly assumed a competitive market in which price P is equal to marginal cost. Moreover, it is assumed that the marginal cost is equal to average cost, but the average cost C is an increasing function of product quality, $C = C(\pi)$ with $C' > 0$. It is the same kind of cost function that Swan [26] and others have used in analyzing the demand for durables; their quality attribute is durability, while mine is the proportion of good units π . Thus, $C(\pi) = P(\pi)$. Let $P' = dP/d\pi$, and $P'' = d^2P/d\pi^2$. We then get

$$\frac{d\sigma}{d\pi} = \left(\frac{1}{P^2}\right) [P(P' + \pi P'') - P'^2].$$

By rearranging terms, it can be shown that $d\sigma/d\pi > 0$ when the following inequality is satisfied:

$$\frac{\pi P''}{P'} > (\sigma - 1).$$

Reference to equation (20) reveals that if $\sigma < 1$, the consumer can reduce the full price P^* by demanding a higher quality product. Thus, the range of admissible product qualities are those for which the price elasticity of quality σ exceeds unity. Given that $\sigma > 1$, the above inequality would be satisfied if $P'' = C'' > 0$ implying that the marginal cost of higher product quality is rising.

³⁴ If there is an upper bound to the distribution of damage costs W , it will fix an upper bound of π that would be produced in a competitive market. The lower bound is set by the point where $\sigma = 1$. Transaction costs, marketing expenses, volume effects, and indivisibilities are likely to limit the breadth of the product line to a finite and probably small number of product grades. To the best of my knowledge, no one has developed a satisfactory theory to explain why burlesque theatres and opera houses offer only a small number of differentiated products, meaning seats of varying qualities and locations. Even local newspapers with some monopoly power establish comparatively simple price structures for their advertisements.

qualities is likely to be narrower under producer liability. Moreover, the full price of good units under producer liability will be higher for those consumers who suffer below average net damage costs thereby placing them on lower indifference curves. When only a small minority of consumers would demand the safer product grades under consumer liability, a shift to strict liability on sellers tends to reduce the potential profitability of safer products.

□ **The home production of safety.** Accidents can be “caused” by a variety of factors: (a) negligence on the part of the consumer-user, (b) defects in the design or manufacture of the product, or (c) simply random chance. If it were possible to design a statistical study that could determine the relative importance of the several contributing factors, we would be in a better position to evaluate the merits of the NCPS recommendations. The Commission recognized that consumer negligence was responsible for many product-related accidents, but they believed that very little could be done to reduce it via making consumers more careful. They concluded that the only way to achieve a significant reduction in the cost of home accidents was through legislated controls over product quality. The issue is thus reduced to an empirical question that can only be resolved by assembling appropriate data which are woefully absent, even in the NCPS report.

The fault system gives the illusion of determining causality by assigning the accident cost to the party judged to be responsible for the accident.³⁵ Surveys of insurance claims and legal suits for damages overstate the importance of product defects. Very few individuals would sue for damages when their own negligence was responsible for the accident. There is considerable evidence suggesting that consumer negligence is a quantitatively important factor contributing to the toll of product-related accidents. Data included in the NCPS report³⁶ reveal, for example, that fully 50 percent of the injuries resulting from the use of rotary lawnmowers involved injury to the hand as some silly operator tried to unclog grass from the mower while the rotary blade was still spinning. Removing safety guards from power tools, placing ladders on boxes, or setting hot water vaporizers on unstable stools offer other examples in which improved product designs would have little impact in reducing the frequency and severity of home accidents. Individuals also differ in their accident proneness; some are clumsier, more reckless, or engage in more hazardous activities. A study by Mellinger et al. offers statistical evidence supporting the thesis of accident proneness as a fairly stable personal trait.³⁷

³⁵ In his article on the fault system, Calabresi [7] makes the obvious but illuminating point that the fault system, by its very nature, minimizes the importance of inherent risks. The case-by-case deliberations under fault try to place the responsibility for an auto accident to driver negligence or to some product defect, even though the auto accident could have been “caused” by a poorly designed road, weather conditions, or a straying cow.

³⁶ In [20], p. 30.

³⁷ The model employed by Mellinger et al. [19] assumed that a Poisson distribution described the probability density of accidents for each individual. If all individuals had the same probability distribution function, a Poisson distribution should describe the observed frequency distribution of accidents per unit time period. But the data (describing the number of medically treated accident injuries of all types for four groups of children over two four-year time intervals) failed to conform to this assumption of equal accident proneness. It was next

The injury toll from product-related accidents could surely be lowered if consumers were prepared to allocate enough time and resources to the home production of safety. They could obtain relevant product information, read warnings, follow instructions, and exercise care in the use of potentially dangerous tools. I agree with the Commission report that some misuse of consumer products is unavoidable. Ladders will be put on boxes, and children will use the hard heads of dolls to discipline their sibs and peers. But the Commission seems to believe that these misuses will continue in spite of efforts at consumer education. The Commission accordingly favored policies that leaned toward placing strict liability on sellers or that controlled quality via federally legislated product standards. A somewhat different view was expressed by McKean who writes:³⁸

The buyer is in a better position than anyone else to know the exact use to which he plans to put a product and what alternative qualities, or degrees of safety, in the product would mean to his costs and gains. The customer, if he is liable, has an extra incentive to acquire and make appropriate use of information. To get the information, he must deal, not with thousands of individuals, but with the seller and a few other identifiable persons.

■ According to the NCPS report, the nation is presently facing a serious product safety problem because consumers are unaware of the risks of using increasingly complex products, unable to cope with these risks even if they knew about them, unable or unwilling to get accurate product information, or misled by questionable advertising practices. In short, the consumer is a pigeon. The atomistic consumer is pitted against a seller who is cast in the role of a powerful and impersonal adversary. He may be forced to consume excessively dangerous products, and he is often unable to buy insurance protection against these product risks at rates near the actuarial odds. These circumstances appear to make the product safety problem similar to that analyzed by Arrow in his article on "Uncertainty and the Welfare Economics of Medical Care."³⁹ In that part of the article which discussed the policy alternatives dealing with uncertainty about product qualities, Arrow concluded:⁴⁰

4. On the policy alternatives before us

The social demand for guaranteed quality can be met in more than one way, however. At least three attitudes can be taken by the state or other social institution toward entry into an occupation or toward the production of commodities in general . . . (1) the occupation can be licensed, non-qualified entrants being simply excluded . . . (2) the state or other agency can certify or label, without compulsory exclusion . . . (3) nothing at all may be done; consumers make their own choices.

assumed that the individual means of Poisson distributions were distributed across persons according to a two-parameter Gamma distribution, and this latter specification provided a far better fit to the data.

When the accident data for the first four-year period were fitted to the model, the children could be classified into low vs. high accident histories implying different means for their respective Poisson distributions. The authors then found that this classification enabled them to predict accident rates for the second four-year period. If there were no accident-prone individuals, or if people learn from their mistakes, the prior accident record should be a poor predictor of future accident rates. Mellinger and his associates found precisely the opposite, suggesting that there are accident-prone persons.

³⁸ In [15], p. 619.

³⁹ In [2].

⁴⁰ *Ibid.*, pp. 966-967.

The choice among these alternatives in any given case depends on the degree of difficulty consumers have in making the choice unaided, and on the consequences of errors of judgment. It is the general social consensus, clearly, that the *laissez-faire* solution for medicine is intolerable.

With respect to product safety, the NCPS report flatly rejects the option of *laissez-faire*. The Commission proposes a combination of the first two options. Under the recommended plan, a Federal agency would be empowered to establish minimum safety standards for existing products. Firms that failed to meet these standards would be excluded from the market. In the case of new products, the agency is charged with the task of compiling accurate and pertinent accident data. If on the basis of these data, the agency concludes that the new product is unreasonably hazardous, it could establish and enforce prescribed safety standards, or it could prohibit the sale of the new product. Are these policies consistent with the goal of maximizing the economic welfare of consumers?

In order to see if similar conclusions can be extended to the product safety problem, it is instructive briefly to review Arrow's model. Arrow constructs the norm of an ideal competitive equilibrium under uncertainty. In this ideal world, the marginal cost of information is zero, and transactions are costless. Risk averse individuals maximize their expected utilities by demanding full insurance coverage against all conceivable risks. In a sense, Arrow has independently reestablished the Coase theorem; namely, with zero information and transaction costs, market forces will lead to an optimal allocation of resources. But information is costly and imperfect, and consumers cannot turn to insurance carriers for all imaginable risks. The moral hazard, administrative costs, and inadequate third-party controls over payment of benefits and costs are some of the reasons identified by Arrow to explain why insurance markets have not been developed for all risks. With respect to medical care, he seems to believe that a nonmarket agency could greatly reduce these insurance costs, and by achieving widespread coverage through compulsion, the agency could eliminate these uncertainty costs, thereby putting risk averse consumers in a preferred position.

What are the implications of adopting Arrow's third policy alternative of *laissez-faire* with respect to product safety? Namely, to do nothing, and to allow consumers to choose among the several qualities of consumer products. Insurance policies for medical care and personal liability that most of us now hold provide us with some protection against the contingent costs of accidents in general. In some cases, consumers are provided with or can purchase service warranties that protect them against product defects. But the availability of insurance specifically designed for consumer product risks is limited for a variety of reasons. It is difficult to verify the legitimacy of claims because most product-related accidents involve only the user, and no one can testify whether the accident was the result of negligence. The transaction costs of insuring risks generated by a multitude of small consumer products (toys, ladders, bikes, snowblowers, etc.) are likely to be so high that at rates consistent with these costs, consumers may choose not to demand insurance coverage.

Calabresi offers several reasons supporting the thesis that private insurance markets typically achieve less than some optimal level

of insurance protection.⁴¹ These include:

- (1) People systematically underestimate the risks, or as Calabresi puts it: "Such things always happen to the other guy, and no amount of statistical information can ever convince the individual that they can happen to him."⁴² A variant of this theme is that even if consumers knew the actuarial risks, they are "psychologically unable to cope with them." This latter variant suggests that the event (be it death or the loss of a limb) is of such an awesome magnitude that the individual is unable to gauge its costs. Underestimation of either the risk or accident cost can thus result in less than optimal insurance protection.
- (2) The paternalism theme argues that because people are short-sighted, inept, or just stupid, society, through the state, must somehow collectively make certain decisions for them. It is alleged that people are happier when someone else makes decisions that force them by law, custom, or contract to set aside part of their incomes to pay for the contingent costs of illness, accident, death, or retirement.⁴³ The implication of this line of reasoning is that people do *not* know what is in their own self interests, but society somehow does and must therefore make the appropriate allocation of resources. It seriously questions the doctrine of consumer sovereignty that the individual is best able to judge for himself what is or is not in his own self interest.
- (3) Some individuals may not buy any insurance because they believe that they will not have to bear the accident costs. Society or the state will pick up the medical bills and subsistence costs in the remote event that they are seriously injured. Such "free riders" thus impose external costs on the rest of us who make up a charitable society that is unwilling to let them suffer the consequences of their improvident behavior .
- (4) Insurance markets are unlikely to be developed when the likelihood of the insured event is significantly affected by the very fact of insurance; i.e. when the "moral hazard" is important.⁴⁴
- (5) Private insurance does not offer sufficient differentiation, and individuals with vastly different accident propensities are often placed in the same risk category. Insurance rates will thus be

⁴¹ In [6], pp. 55–64. Although Calabresi agrees with the conclusion that private insurance alone leads to less than a socially optimal level of insurance, he does *not* concur with all of the reasons set forth in his book. A fuller discussion of my evaluation of the merits of these arguments can be found in [24], pp. 98–102.

⁴² *Ibid* , p. 56.

⁴³ A closely related theme is what Calabresi calls the "Faust attitude" that many individuals are myopic when it comes to choices between immediate costs and expected future long-range costs. As he describes it, "In such cases, people tend to choose the immediate good life only to regret it later" [6], p. 57. In the economist's jargon, the individual with the Faust attitude has a very high time preference and is thus willing to substitute a very high future cost for a small reduction in current cost. However, in retrospect, he regrets his choices in the past, presumably because his retrospective time preference rate is different. (Some of us on the morning after the carousal can surely sympathize with this attitude.)

⁴⁴ Pauly [25] discusses the implications of moral hazard for medical care insurance. Arrow [1] refers to the phenomenon as "the confounding of risks and decisions." Thus, one cannot insure his house for more than its market value because the insurance policy may give the insured an incentive to burn down his house.

too high for some and too low for others thereby leading to nonoptimal demands.

Given the broad range of potentially hazardous products and the wide diversity in accident propensities, it seems unlikely that private insurance carriers would offer policies expressly designed for product-related accidents.

The poor consumer who lacks the information to make a "rational" choice is presumed to err in the direction of shoddier and riskier products. We rarely hear about the other errors where consumers enjoy windfalls from a good buy of a safe product. It has been argued that consumers cannot get enough product information because sellers lack the incentives to produce such information. To the extent that there are external benefits from the production of better information, there is a basis for intervention, possibly via the government subsidization of information production.⁴⁵ This last suggestion comes close to Arrow's second policy alternative, namely certify or label without compulsory exclusion. It should be noted, however, that the NCPS position on the social value of better information and consumer education is at odds with Arrow's analysis. The Commission felt that the accident toll could not be appreciably lowered by supplying consumers with more information.

The policies recommended in the NCPS report are intended to protect the consumer by establishing barriers to the entry of unreasonably hazardous products. Product standards would have to be specified by some agency for a wide range of products. Tests would have to be performed to determine if a new entrant in an established product line met these standards. The task of setting these standards would be a costly one, especially for a new product because of the varying environments in which the product will be used and the different accident propensities of potential users. Something akin to the bureaucracy of the Food and Drug Administration would probably emerge to license the admissible dolls, bikes, window glass, cleansers, and power mowers that consumers will be permitted to buy.⁴⁶

In closing, let us review the conclusions of the NCPS report. The nation presently incurs a substantial cost as the result of product-related accidents. Some of these accidents are undoubtedly caused by unreasonably hazardous products. The Commission concluded that the post-accident remedies available to consumers in the form of compensation for damages were judged to be inadequate and encouraged manufacturers to continue the production and sale of un-

⁴⁵ Some institutions like *Consumer Report* and various trade journals have developed to provide consumers with product information. In addition to these, one can turn to various informal channels. An established department store or other retail outlet will typically provide its customers with implied product warranties, and will usually not stock the most hazardous designs of certain products. The question of "What is a socially optimal quantity of information?" is an extremely difficult one that is beyond the scope of this paper. The interested reader is referred to Arrow [1] and Hirshleifer [13].

⁴⁶ In a recent article [17] Professor McKean analyzes the difficult problems of regulating chemicals and drugs. He mentions, for example, that a new oral contraceptive would now require nearly fourteen years of testing before it would be approved for sale. Moreover, much of the testing now required for licensing a new drug is of questionable value.

safe products.⁴⁷ It was further concluded that competitive market forces, better product information, and industry self-regulation would fail to provide consumers with “enough safety.” They accordingly recommended that sellers be made strictly liable for accident costs. But I have demonstrated in this paper that a move to strict liability (meaning that all sellers would be forced to offer the tied package of the risky product plus a full coverage insurance policy) need not always lead to the increased production of safer goods. Moreover, strict liability on sellers would redistribute wealth in favor of those consumers who incur high accident costs (and who presumably are wealthier) and against poorer consumers who prefer riskier products at lower prices. In the worst case, in which the mean damage cost $\bar{W}$ is less than the critical damage cost W^* , we not only redistribute wealth away from the poor, but also increase the aggregate cost of accidents.

However, the stronger policy recommended by the NCPS would establish direct government controls that could exclude “unreasonably hazardous products” from the market. The goal of public policy toward product safety should not be the minimization of accident costs. If that were the goal, we should not permit any rotary lawnmowers but insist instead on the old-fashioned unpowered mowers or as a second-best, allow only reel-type power mowers since these restrictions would reduce home accidents. The appropriate goal should be the maximization of the economic welfare of consumers who, when given a choice, will choose that product grade which minimizes the *sum* of accident costs and accident prevention costs. Policies that give consumers greater freedom of choice and that place less reliance on direct government controls would seem to come closer to this goal.⁴⁸

References

1. ARROW, K. J. “The Organization of Economic Activity: Issues Pertinent to the Choice of Markets vs. Nonmarket Allocation,” in *Joint Economic Com-*

⁴⁷ Even in the absence of product warranties, a consumer can always sue, but the NCPS argued that this alternative was ruled out for most consumers by the high legal costs of court actions. A slightly different impression is given in a recent article, “Buyer vs. Seller in Small Claims Courts” [10]. A survey conducted by *Consumer Report* revealed that in the vast majority of cases in which the consumer was the plaintiff (in suits for damages from defective products, failure to perform, breach of warranty, etc.), he felt that the courts were fair and equitable. (However, the consumer felt the courts were unfair in cases where he was the defendant, usually involving the collection of bad debts.) The *Consumer Report* study felt that the present upper limits for damages in most states (which allow the plaintiff to use the Small Claims Courts) were inordinately low. Thus, in some states, cases involving damage claims of as little as \$250 were denied access to Small Claims Courts where the plaintiff does not have to be represented by legal counsel. They recommended that the upper limit be raised to \$1000 which seems quite reasonable to me.

⁴⁸ What appears to be an unreasonably dangerous product to a middle-aged cautious, risk averse person may be a wholly acceptable risk to a young, risk-preferring man. Thus, in Section 3, it was shown that each consumer would demand that product quality which minimized the sum of the accident prevention cost (in terms of a higher price for safer grades) plus the expected accident cost, $W\gamma$. There is, indeed, considerable evidence suggesting that the attitude toward risk-bearing as well as the accident cost are systematically related to the age and sex of the individual; see Oi [24], pp. 45–75. Finally, it should be remembered that the Federal Trade Commission already has the power to prosecute sellers who engage in fraudulent advertising.

- mittee, *The Analysis and Evaluation of Public Expenditures: The PPB System*, Vol. I, Washington, D. C.: U. S. Government Printing Office, 1969, pp. 47-64.
2. ———. "Uncertainty and the Welfare Economics of Medical Care." *The American Economic Review*, Vol. 53, No. 5 (December 1963), pp. 941-973.
3. ASHER, N. S., BEAZER, W. F., COX, W. A., MUTH, R. F., AND OI, W. Y. *Demand Analysis for Air Travel by Super-sonic Transport*. Report No. R118, Institute for Defense Analysis, Arlington, Va., December 1966, in two volumes.
4. BECKER, G. S. "A Theory of the Allocation of Time." *Economic Journal*, Vol. 74, No. 295 (September 1965), pp. 493-517.
5. BUCHANAN, J. M. "In Defense of *Caveat Emptor*." *University of Chicago Law Review*, Vol. 38 (Fall, 1970), pp. 64-73.
6. CALABRESI, G. *The Costs of Accidents: A Legal and Economic Analysis*. New Haven: Yale Univ. Press, 1970.
7. ———. "Does the Fault System Optimally Control Primary Accident Cost?" *Law and Contemporary Problems*, Vol. 33, No. 3 (Summer 1968), pp. 439-463.
8. ——— AND BASS III, K. C. "Right Approach, Wrong Implications: A Critique of McKean on Products Liability." *University of Chicago Law Review*, Vol. 38 (Fall 1970), pp. 74-91.
9. COASE, R. H. "The Problem of Social Cost." *Journal of Law and Economics*, Vol. 3, No. 2 (October 1960), pp. 1-44.
10. CONSUMER REPORT. "Buyer vs. Seller in Small Claims Courts." *Consumer Report*, Vol. 36, No. 10 (October 1971), pp. 624-631.
11. HEFLEBOWER, R. H. "The Theory and Effects of Nonprice Competition." in R. E. Kuenne, ed., *Monopolistic Competition Theory: Studies in Impact*. New York: John Wiley and Sons, Inc., 1967, pp. 177-201.
12. HILLIER, F. S. AND LIEBERMAN, G. J. *Introduction to Operations Research*. San Francisco: Holden-Day, Inc., 1967.
13. HIRSHLEIFER, J. "The Private and Social Value of Information and the Reward to Inventive Activity." *The American Economic Review*, Vol. 61, No. 4 (September 1971), pp. 561-574.
14. LANCASTER, K. J. "A New Approach to Consumer Theory." *Journal of Political Economy*, Vol. 74, No. 2 (March-April 1966), pp. 132-157.
15. MCKEAN, R. N. "Products Liability: Implications of Some Changing Property Rights." *Quarterly Journal of Economics*, Vol. 84, No. 4 (November 1970), pp. 611-626.
16. ———. "Products Liability: Trends and Implications." *University of Chicago Law Review*, Vol. 38 (Fall 1970), pp. 3-63.
17. ———. "Property Rights, Regulation of Chemicals, and Information Production." Unpublished manuscript, University of Virginia, 1972.
18. MANNE, H. G. "Edited Transcript of AALS-AEA Conference on Products Liability." *University of Chicago Law Review*, Vol. 38 (Fall 1970), pp. 117-141.
19. MELLINGER, G. D., SYLVESTER, D. L., GAFFEY, W. R. AND MANHEIMER, D. I. "A Mathematical Model with Applications to a Study of Accident Repeatedness among Children." *Journal of the American Statistical Association*, Vol. 60, No. 312 (December 1965), pp. 1046-1059.
20. NATIONAL COMMISSION ON PRODUCT SAFETY. *Final Report of the National Commission on Product Safety*. Washington, D. C.: U. S. Government Printing Office, June 1970.
21. ———. *Industry Self-regulation, Supplemental Studies*. Vol. 2. Washington, D.C.: U. S. Government Printing Office, June 1970.
22. ———. *Product and Injury Identification, Supplemental Studies*. Vol. 1. Washington, D.C.: U. S. Government Printing Office, June 1970.
23. ———. *Product Safety, Law, and Administration, Supplemental Studies*. Volume 3. Washington, D. C.: U. S. Government Printing Office, June 1970.
24. OI, W. Y. "An Essay on Workmen's Compensation and Industrial Safety." Unpublished manuscript, University of Rochester, 1972.
25. PAULY, M. V. "The Economics of Moral Hazard: Comment." *The American Economic Review*, Vol. 58, No. 3 (June 1968), pp. 531-537.
26. SPENGLER, J. J. "The Economics of Safety." *Law and Contemporary Problems*, Vol. 33, No. 3 (Summer 1968), pp. 619-638.
27. SWAN, P. L. "Durability of Consumption Goods." *The American Economic Review*, Vol. 60, No. 5. (December 1970), pp. 884-894.

An econometric planning model for American Telephone and Telegraph Company

B. E. Davis,
G. J. Caccappolo,
and
M. A. Chaudry

Corporate Planning Organization
American Telephone and Telegraph Company

In this paper an econometric model for corporate planning analysis is discussed and its operation is demonstrated by examples. The thrust of this effort has not been towards developing a new body of theory or detailed econometric examination of standard relationships in the theory. The main thrust has been towards the development of a capability for analyzing firm behavior utilizing economics in a straightforward fashion. The structure and econometrics simply allow the theory to be examined quantitatively. In a sense, this has been the theme of this development; that is, for present day corporate planning modeling, the state of applied economics and the availability of planning technology is such that comprehensive and informative models can be constructed on simple economic theory.

■ The development of models of economies is well into its second generation.¹ Sector models or models of industries within an economy are following in a satellite fashion² and most recently as separate model developments.³ The objective of this paper is to report on a corporate planning model development which explicitly attempts to interrelate the economics of demand, production, and finance in a framework which accounts for the impact of the national economy

1. Introduction

The model development reported in this paper is a result of support by many groups and organizations. These include individuals within the Corporate Planning Organization, selected areas in the General Departments of A.T.&T., and the professional staff of Data Resources, Inc. The authors are especially grateful to Professor P. W. MacAvoy of M.I.T. for his interaction and consultation, to Dr. O. Eckstein of Harvard for his suggestions and encouragement, and to R. L. Breedlove of A.T.&T. whose support throughout was central in the development of the model. A portion of this work has been reported earlier in *Corporate Planning Organization Report*, A.T.&T. Co. [6].

Blaine E. Davis received the B. S. degree in electrical engineering from the University of Maine at Orono (1964). He holds the M.S. in operations research (1965) and the Ph.D. in economics and operations research (1969) from The Johns Hopkins University. Dr. Davis is Manager—Economic and Demand Studies in the A.T.&T. Company's Corporate Planning Organization. His current research activities are concerned with econometric corporate modeling for planning purposes with regard to public policy.

Gerard J. Caccappolo received the B.S. degree in electrical engineering from Manhattan College (1963), the M.S. degree in engineering from N.Y.U. (1965), and the M.S. degree in operations research from Columbia University (1970). Mr. Caccappolo is Supervisor—Economic Analysis in the Corporate Planning



and other environmental (regulatory) factors. The thrust of the effort has not been towards developing new theory or detailed econometric examination of existing relationships. As a matter of fact, in a sense, the opposite has been true in that standard economic theory has been used and applied in a straightforward fashion.⁴ The main thrust is towards a capability for analyzing firm behavior.

The paper is divided into four major parts. First will be an overview discussion of the structural form of the model. Next, each sector of the model will be examined in detail including specification of the econometric relationships. Following this, the functional operation of the model will be reviewed in one of its operational modes including an examination of its 1972 performance. Finally, two examples of policy analysis utilizing the model will be presented.

2. General features of the model

■ The econometric model of the Bell System has been developed as a planning tool to assist in evaluating the impact of general economy changes and other environmental changes on alternative Bell System policies. These policies are concerned with long-range planning. The model is called FORECYT, an acronym for Econometric FORecast-ing Model for Corporate PoliCY Analysis for T (the ticker symbol for A.T.&T.). The overall modeling approach is based on the premise that the state of the economy determines an individual firm's demand, making it externally derived rather than created by the firm's supply capability. With this mode dominating, Bell System demand is assumed dependent upon economic factors external to the Corporation and supply is a reaction, via corporate policy actions, to the demand.

The main driving element for the model is a forecast of demand from the forecasted state of the U. S. economy. This "state of demand" is a central function of the model, in conjunction with fore-

Organization of A.T.&T., where his present activities include research into investment decision theory and technological innovation.

Mohammad Ali Chaudry received the B.Sc. in economics from The London School of Economics and Political Science (1967). He was awarded the M.A. and the Ph.D. in economics by Tufts University (1969 and 1971, respectively). Dr. Chaudry is Supervisor—Economic Studies in A.T.&T.'s Corporate Planning Organization. His research interests include econometric demand analysis, micro-econometric models, and the role of telecommunications in development and planning.

¹ If one is willing to label the Wharton Model of the U. S. Economy [18] as the first generation of models for evaluation and planning at the macroeconomic level, then many of the more recent developments of econometric models of the U. S. Economy such as the Penn-M.I.T.-F.R.B. model and that of O. Eckstein at Data Resources, Inc., could be called second generation.

² Most of the recent commercially available macro models such as that of Chase Econometrics, Inc., and Data Resources, Inc., have satellite sector models of varying comprehensiveness.

³ The reference here is to the recent model development reported by R. Dobell et al. [7] where a model of telecommunications in Canada was developed. In addition, Dobell developed a number of empirical relationships for Bell Canada in a manner analogous to the Canadian sector analysis. With additional modeling effort and with the framework to be presented in this paper, the Bell Canada Analysis could be structured into a corporate model similar to the development here.

⁴ One aspect of this development was to isolate, in a systems sense, where the weakest link in the "chain" of modeling was. For this application it was not in the traditional areas of demand, production, or financial modeling. This topic is discussed in detail later.

cast price changes due to anticipated regulatory and Federal Cost of Living Council actions. In addition, the posture of the financial world in which the Corporation operates is controlled, at least in part, by the general conditions of the economy. Collectively, these influences determine not only the forecasts of Bell System demand but also the costs which will be incurred and the revenues which will accrue from alternative modes of response.

FORECYT has a tripartite structure (see Figure 1) consisting of three submodels, which are further divided into modules to facilitate disaggregation as interest dictates and data allow. The three submodels are:

- (1) The *Environment Model*, a construct of the "world" or environment with which the Corporation interacts, which contains a National Economy Module, a Regulatory Module, and a Price-Wage Review Module.
- (2) The *Corporate Model*, an econometric construct of the Bell System which contains a Demand Module, a Price Module, a Capital Market Module, a Financial Module, a Revenue Module, and a Production (or Supply) Module.
- (3) The *Management Model*, a logical construct providing policy variables for control (in the Input Module) and displaying corporate indicators of performance (in the Output Module).

■ This section describes the behavioral structure of each submodel of FORECYT in terms of the behavioral equations and interacting relationships of the module blocks. The schematic outline of FORECYT shown in Figure 2 is implicitly assumed in this description and should be used as a structural guide.

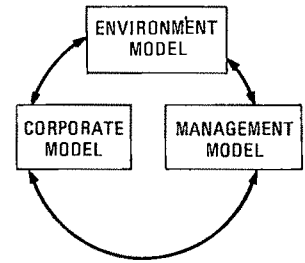
□ **Environment model.** The Environment Model contains a National Economy Module which provides the forecasted state of the national economy, a Regulatory Module, which depicts a simple single regulatory commission for the Bell System,⁵ and a Price-Wage Review Module, which reflects the actions of the Federal Cost of Living Council.

National economy module

Of all the modules in the Environment Model, the National Economy Module (see Figure 3A) is by far the most complex. It is *not* a model built or maintained within the FORECYT structure. Its essential role is to provide forecasts of the U. S. economy in terms of quantities and variables which influence the Corporate Model. For conceptual consistency, this module needs to be an econometric model of the U. S. economy which reflects interval by interval its anticipated state in terms of forecasts of influential variables. There are a number of macroeconomic models of the U. S. economy which can and do provide these variables. FORECYT has been designed to use any one of these, or all of them sequentially, to analyze

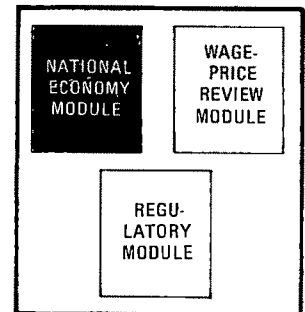
⁵ This is a significant abstraction from reality since the Bell System deals with in excess of fifty separate regulatory commissions. Detail in this area will develop as more disaggregate versions of the model are developed.

FIGURE 1
TRIPARTITE STRUCTURE
OF FORECYT



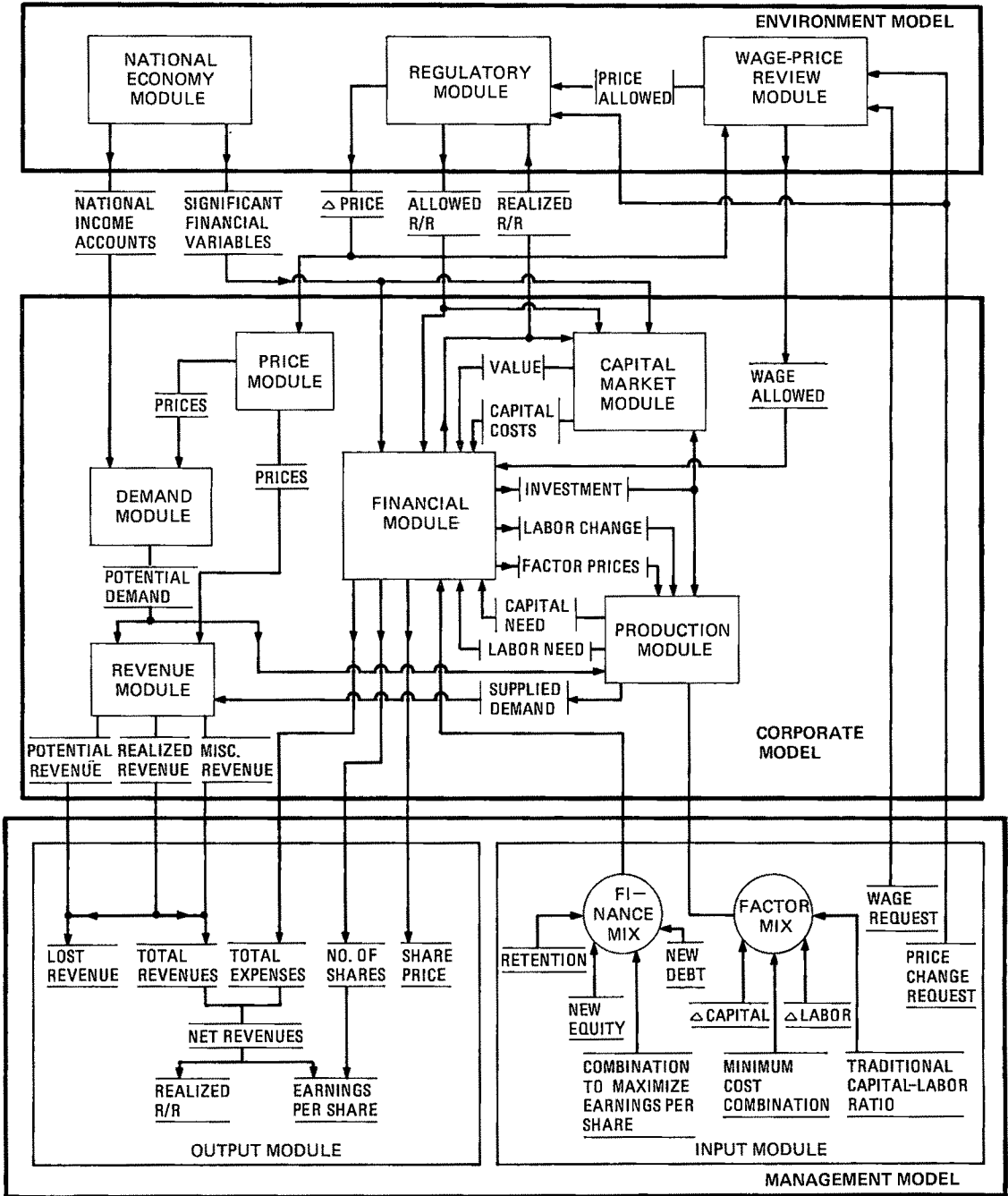
3. Description of model structure

FIGURE 3A
ENVIRONMENT MODEL



AN ECONOMETRIC
MODEL OF AT&T / 31

FIGURE 2
FORECYT



the impact, if any, of differing views of the forecasted state of the U. S. economy.

Wage-Price Review module

To investigate the possible effects of “double price regulation,” a module reflecting a range of possible actions stemming from the Federal Cost of Living Council has been included in FORECYT (see Figure 3B). In its present form, the quantitative mechanism reflecting this activity is defined in terms of constraints on maximum wage and

price increases. These are represented in the module as

$$\frac{W(t) - W(t-1)}{W(t-1)} \leq \text{Wage rate increase guideline}$$

and

$$\frac{P_i(t) - P_i(t-1)}{P_i(t-1)} \leq \text{Price change guideline,}$$

where

$W(t)$ = Wage rate at time t , and

$P_i(t)$ = Price of service i at time t .

(Price indices are used in the model.)

The wage-price review takes place principally in response to requests for wage or price changes by the Management Module.⁶

Regulatory module

The current version of the Regulatory Module (see Figure 3C) is an aggregate representation of the regulatory process facing the corporation. Its primary function is to induce price changes into the Corporate Model. An analytical representation of a continuous surveillance feature of this module is outlined in the statements that follow.

Realized rate of return is defined as

Rate of Return

$$= \frac{(\text{Total Revenues}) - (\text{Total Expenses other than Capital})}{\text{Rate Base}}$$

The economic representation of this definition can be expressed as

$$r_t = \frac{P_t Q_t - (WL)_t}{K_t}, \quad (1)$$

where

r_t = Realized rate of return at time t ,

P_t = "Price" at time t ,

Q_t = "Quantity" at time t ,

$(WL)_t$ = Expenses (labor) at time t , and

K_t = Capital at time t .

Assuming a continuous-surveillance regulatory form of "test year's output at the following year's prices to determine prices for desired rate of return" yields

$$r_t^* = \frac{(P_t + \Delta P_t) Q_t - (WL)_t}{K_t}, \quad (2)$$

⁶ There is also a price change "loop" from the Regulatory Module to allow analysis of situations where the Regulatory Module would prescribe a price change in excess of the guidelines allowed by the Federal Cost of Living Council. Another option allows both wage and price changes to feed through this module unconstrained. Finally, the guidelines for wage and price increases can easily be changed to analyze the effects of possible future phases of this type of regulation (i.e., a possible "Phase 4").

FIGURE 3B

ENVIRONMENT MODEL

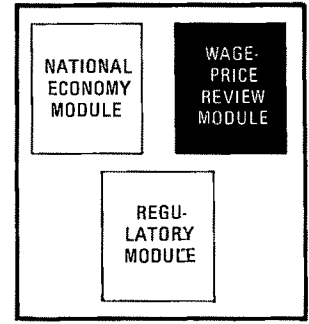
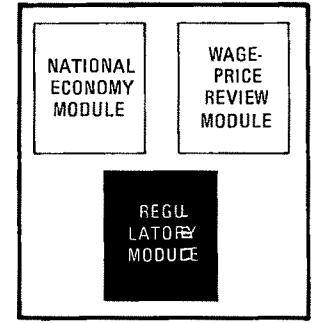


FIGURE 3C

ENVIRONMENT MODEL



which can be written as

$$\Delta P_t = (r_t^* - r_t) \frac{K_t}{Q_t},$$

where

ΔP_t = Required price change at time t , and

r_t^* = Required (or desired) rate of return.

Two additional modifications to the price-change equation are desirable. First, the allowed (or desired) rate of return can be related to the cost of capital by

$$r_t^* = \alpha \frac{1}{N} \sum_{j=1}^N \rho_{t-N+j}, \quad (3)$$

where

ρ_t = Cost of capital at time t ,

α = Proportionality factor, and

N = Number of periods over which cost of capital is averaged.

If both α and N are equal to one, then the current allowed rate of return is defined as being equal to the current-period cost of capital, and would then be used to determine the next year's prices.

The second change is related to the lag between the authorized price changes and the actual changes instituted by the company. The size of the differential between realized and allowed price will in part be dependent upon the responsiveness of the regulatory process. A measure of this can be included in the price change equation as

$$\Delta P_t = \beta \left[(r_t^* - r_t) \frac{K_t}{Q_t} \right], \quad (4)$$

where

$$\frac{1}{\beta} = \text{lag period between authorized and instituted price changes.}^7$$

The current version of FORECYT has been tested in the mode where price changes are dictated exogenously. Data are currently being collected to estimate the continuous-surveillance equation statistically with regulatory lag.

□ **Corporate model.** The Corporate Model is an econometric representation of the Bell System. It is the central submodel of FORECYT, providing the mechanisms which show how crucial variables from national economic forecasts and proposed corporate policies interact. Comprising the Corporate Model are econometric modules which represent various areas of Bell System operation. There is a Price Module, a Demand Module, a Revenue Module, a Production Module, a Financial Module, and a Capital Market Module. In the construction of these modules, emphasis has been equally concentrated on empirical robustness and theoretical consistency. It is of

⁷ A small " β " means a long lag period, where prices are unresponsive to differences between allowed and realized rate of return. If β were statistically estimated from historical data, its range would be expected to be

$$0 < \beta < 1.$$

central importance to forecast the consequences of proposed courses of action; however, from a policy viewpoint, it is of equal importance to know why these outcomes are anticipated and how they may change under different courses of action.

Price module

The principal function of the Price Module (see Figure 4A) is to calculate appropriate price indices and the mechanisms for changing these indices, as price changes come from the Regulatory Module. For each service, a Laspeyres Index of the base-weighted aggregative form has been used. Within each class of service, the prices of different types of calls have been aggregated in a weighted sum from the base year 1970. The following four price indices have been used in estimating the demand equations constituting the Demand Module of FORECYT, described below:

PLOC = Price index for Local Telephone Service,

PTOL = Weighted price index for Interstate and Intrastate Message Toll Service,

PWATS = Price index for Outward WATS (Interstate), and

PPL = Price index for Private Line Service (Interstate).

A brief description of each index is given in the Appendix.

Demand module

The demand for telecommunications services is determined by the state of the national economy, demography, prices, and consumer tastes. The Demand Module of FORECYT (see Figure 4B) represents these relationships for specified classes of services and their economic determinates. To avoid unit specification and service aggregation problems, demand is measured by a surrogate value index measure of output for each service. This surrogate is derived from revenues as follows:

$$\frac{R_{it}}{P_{it}} = Q_{it} = \text{the measure of quantity demanded of service } i \text{ at time } t,$$

where

R_{it} = Revenue from service i at time t , and

P_{it} = Price Index of service i at time t .

The specification of Q_{it} follows standard economic theory in that

$$Q_{it} = Q_{it} \left(\frac{\text{Price}_{it}}{P_t^*}, \frac{\text{Income}}{P_t^*}, \text{Habit}, \text{Market} \right),$$

where

P_t^* = General price deflator (e.g., implicit GNP deflator).

The functional form of Q_{it} is not specified by theory; the form with the best results was a product form of the demand function yielding

$$\begin{aligned} \log(Q_{it}) = & \alpha_i + \beta_i \log(Q_{i,t-1}) + \gamma_i \log\left(\frac{Y_{it}}{P_t^*}\right) \\ & + \delta_i \log\left(\frac{P_{it}}{P_t^*}\right) + \eta_i \log(Z_{it}), \end{aligned} \quad (5)$$

FIGURE 4A
CORPORATE MODEL

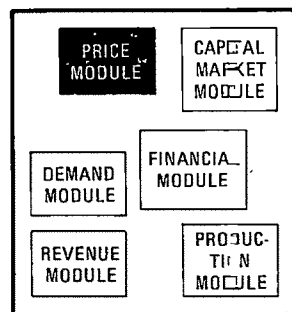
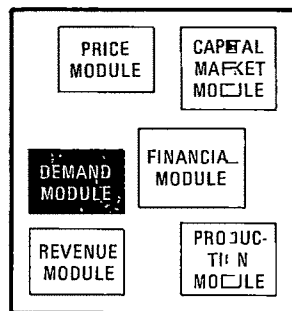


FIGURE 4B
CORPORATE MODEL



where

- $Q_{i,t-1}$ = Habit variable,⁸
 Y_t = Income level variable,
 P_t = Price index, and
 Z_t = Market potential variable.

Standard economic theory requires the signs and magnitudes of the regression coefficients to be

- $\gamma_i > 0$ Income elasticity of demand for service i ,
 $\delta_i < 0$ Price elasticity of demand for service i ,
 $\gamma_i + \delta_i = 0$ Consumer demand functions homogeneous of degree zero in income and price,
 $0 < \beta_i < 1$ Habit variable chosen as last period's consumption,
 $\eta_i > 0$ Positive influence of market size, and
 α_i Unspecified.

Demand equations for Local Service (excluding Local Private Line), Message Toll and other toll services, Wide Area Telephone Service, and Private Line Service have been estimated using the above form.

The econometric demand equations are given below including summary statistics (quantities in parentheses are standard errors of the coefficients). Ordinary least squares estimates have been used throughout the analysis.

Local Service— Q_{it}

$$\log(Q_{it}) = -3.0092 + 0.3351 \log(Q_{it-1}) - 0.2076 \log\left(\frac{PLOC}{PGNP}\right) + 0.2499 \log\left(\frac{I}{PC}\right) + 0.6823 \log(TT) \quad (6)$$

(1.4189) (0.1418) (0.0991) (0.0572) (0.1742)

$$\bar{R}^2 = 0.9982 \quad D.W. = 2.13 \quad SER = 0.0074,$$

where

Q_{it} = Local Service revenues, excluding Private Line, deflated by the local price index,

$PLOC$ = Local Service price index,

$PGNP$ = Implicit GNP deflator,

I = Personal disposable income per capita (16 years and over) less government transfers to persons,

PC = Implicit deflator for personal consumption, and

TT = Total telephones excluding residential extensions.

The period of fit is from 1961:2 to 1971:4, quarterly.

⁸ The past level of consumption, $Q_{i,t-1}$ is used as a proxy for tastes and habit formation. Unlike consumer durables, consumption of services such as communications has a strong element of habit formation. This influence seems quite independent of current income and relative prices. While this choice of a habit surrogate variable works well for the short run, there are some interesting features

Message Toll Service— $Q_{t,t}$

$$\log(Q_{t,t}) = -8.2432 + 0.1395 \log(Q_{t,t-1}) - 0.8847 \log\left(\frac{PTOL}{PGNP}\right) \\ + 0.8287 \log\left(\frac{I}{PC}\right) + 1.4755 \log\left(\frac{TT}{N}\right) \quad (7)$$

(0.57069) (0.1364) (0.1763) (0.1377) (0.4800)

$$\bar{R}^2 = 0.9978 \quad D.W. = 2.08 \quad SER = 0.0147,$$

where

Q_t = Message Toll revenues, including Other Toll, deflated by the toll price index,

$PTOL$ = Message Toll price index,

$PGNP$ = Implicit GNP deflator,

I = Personal disposable income per capita (16 years and over) less government transfers to persons,

PC = Implicit deflator for personal consumption,

TT = Total telephones, excluding residential extensions, and

N = Total population 16 years and over.

The period of fit is from 1962:2 to 1971:4, quarterly.

In each of the above demand equations total telephones excluding residential extensions is used as an explanatory variable and needs to be estimated. The equation is:

Total Telephones— TT

$$\log\left(\frac{TT}{N}\right)_t = 2.7170 + 0.7807 \log\left(\frac{TT}{N}\right)_{t-1} \\ - 0.0167 \log\left(\frac{PLOC}{PGNP}\right) + 0.0846 \log\left(\frac{GNP}{N}\right) - 0.0035 \text{ STRIKE} \quad (8)$$

(0.5895) (0.0417) (0.0234) (0.0153) (0.0009)

$$\bar{R}^2 = 0.9995 \quad D.W. = 2.00 \quad SER = 0.0015,$$

where

TT = Total telephones, excluding residential extensions,

GNP = Gross National Product,

$PLOC$ = Local Telephone Service price index,

$PGNP$ = Implicit GNP deflator,

N = Population 16 years and over, and

$STRIKE$ = Dummy variable for strikes (1.0 for 1968:2, 3, and 1971:2, 3, 4; zero otherwise).

The period of fit is from 1961:1 to 1971:4, quarterly.

associated with this choice for the long run. In particular its effect on long-run price elasticity. While the model specification has price elasticity as a constant, δ , the long-run elasticity can be shown to depend on the parameter, β , the coefficient of the habit variable. R. L. Shanahan has examined the effects of this choice and the consequences of alternative surrogates [15].

Wide Area Telephone Service (WATS)— $Q_{w,t}$

$$\log(Q_{w,t}) = 1.9195 + 0.8289 \log(Q_{w,t-1}) - 0.1434 \log\left(\frac{PWATS}{PGNP}\right) + 0.1785 \log(S451DNS) \quad (9)$$

(0.1427) (0.0131) (0.0534) (0.0303)

$$\bar{R}^2 = 0.9993 \quad D.W. = 2.36 \quad SER = 0.0176,$$

where

Q_w = WATS revenues deflated by the WATS price index,

$PGNP$ = Implicit GNP deflator, and

$S451DNS$ = Sales of domestic airlines, not seasonally adjusted.

The period of fit is from 1962:1 to 1971:2, quarterly.

Private Line Service— $Q_{P,t}$

$$\log(Q_{P,t}) = -6.6581 + 0.1813 \log(Q_{P,t-1}) - 0.7377 \log\left(\frac{PPL}{PGNP}\right) + 1.2813 \log(TPLT) \quad (10)$$

(1.2122) (0.0978) (0.0895) (0.1705)

$$\bar{R}^2 = 0.9872 \quad D.W. = 1.18 \quad SER = 0.0307,$$

where

Q_P = Private Line revenues, local and toll, deflated by the Private Line price index,

PPL = Private Line price index,

$PGNP$ = Implicit GNP deflator, and

$TPLT$ = Total Private Line telephones.

The period of fit is from 1961:2 to 1971:4, quarterly.

Private Line Service demand has total Private Line telephones as an explanatory variable and it needs to be estimated. The equation is:

Private Line Telephones— $TPLT$

$$\log(TPLT)_t = 0.0499 + 0.9999 \log(TPLT)_{t-1} - 0.0086 \log\left(\frac{PPL}{PGNP}\right) \quad (11)$$

(0.1030) (0.0064) (0.0083)

$$\bar{R}^2 = 0.9987 \quad D.W. = 1.71 \quad SER = 0.0049.$$

TABLE 1

ELASTICITIES (SHORT -
RUN ESTIMATES)

SERVICE	PRICE	INCOME
LOCAL	-0.2076	0.2499
TOLL	-0.8847	0.8287
PRIVATE LINE	-0.7377	-
WATS	-0.1434	-

The period of fit is from 1961:1 to 1971:4, quarterly.

Comparisons between the services in terms of price and income elasticities shown in Table 1 are indicative of the characteristics of these services.⁹ For local and message toll, the toll service is seen to be more sensitive to price changes than local. However, each has an

⁹ Comparisons between these estimates and estimates of other investigators are difficult since definition of price and service vary widely. Perhaps the closest comparison can be made to the work of Dobell et al. [7], however, here also, there are definitional differences (in particular in the price indices).

elasticity less than unity which would be expected under regulation. In addition, the relationship between price and income exhibits the homogeneity of degree zero condition.

The Private Line and WATS equations reflect the particular market characteristics of each service. For WATS the principle determinate is viewed as the service offering itself, hence price is quite inelastic (compared to alternatives). The variable $S451DNS$ is indicative of this stylized market structure. The Private Line equation displays reasonable price sensitivity but habit formation was not found to be significant. The coefficient for $TPLT$ is greater than unity, indicating that only a part of private line revenues comes from terminals measured in terms of number of phones connected. A large share of revenues comes from interexchange channels leased by customers.

Revenue module

Total Revenues are the sum of Local, Toll, WATS, and Private Line Revenues plus Miscellaneous Revenues less Uncollectible Revenues. The Revenue Module (see Figure 4C) is defined as

$$\text{Total Revenues at time } t = \sum_i P_{it} Q_{it} + \text{Miscellaneous Revenues} - \text{Uncollectibles}, \quad (12)$$

where

i = Local, Toll, WATS, and Private Line Services.

Miscellaneous Revenues were viewed as being strongly related to the levels of the principal demands on the network.

Miscellaneous Revenues—MREV

$$\begin{aligned} (MREV)_t = & 7690.06 + 0.7473 (MREV)_{t-1} + 0.0105 [PLOC \cdot Q_t \\ & (3186.91) \quad (0.0938) \quad (0.0035) \\ & + PTOL \cdot Q_t + PWATS \cdot Q_w + PPL \cdot Q_P]_t \quad (13) \\ \bar{R}^2 = & 0.9989 \quad D.W. = 2.54 \quad SER = 888.93. \end{aligned}$$

The period of fit is from 1960:2 to 1970:4, quarterly.

Uncollectible Revenues were viewed to be principally a function of the level of Local and Toll Revenues. Private Line and WATS Services were not considered to be significant contributors.

Uncollectible Revenues—UREV

$$\begin{aligned} \log (UREV)_t = & -4.9934 + 0.7612 \log (UREV)_{t-1} \\ & (2.0169) \quad (0.1189) \\ & + 0.4951 \log [PLOC \cdot Q_t + PTOL \cdot Q_t] \quad (14) \\ & (0.2097) \\ \bar{R}^2 = & 0.9735 \quad D.W. = 1.53 \quad SER = 0.0577. \end{aligned}$$

The period of fit is from 1962:1 to 1970:2, quarterly.

Production module

The principal function of the Production Module (Figure 4D) is to determine the levels of the two primary factors of production (labor and capital) needed to meet demand.

FIGURE 4C
CORPORATE MODEL

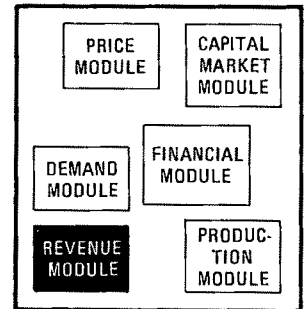
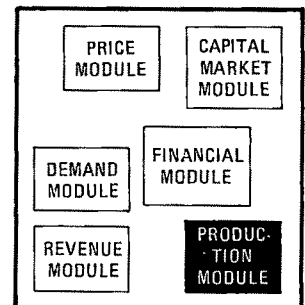


FIGURE 4D
CORPORATE MODEL



An aggregate Bell System production function was generated using a capital stock series and labor series constructed from Bell System corporate data. The capital stock series was constructed along the line of Kuh's measurement of capital stock,¹⁰ i.e., cumulated gross investment less retirements, each deflated by an appropriate index. This measurement of capital stock assumes no loss of efficiency over time of existing capital.

The gross investment series consists of Bell System construction program expenditures (*CP*). The deflator used is a macro capital stock price deflator for private fixed nonresidential investment (*PIFIXER*).

In summary, the capital stock series (*K*) is constructed from the following relationship:¹¹

$$K(t+1) = K(t) + \frac{CP(t)}{PIFIXER(t)} - \frac{\text{Retirements}(t)}{PIFIXER(t-23)} \quad (15)$$

In using this capital stock series to estimate the production function, an underlying assumption of constant plant utilization over the period of estimation was made¹² so that

$$\begin{aligned} K'(t) &= \mu K(t), \\ \text{services} & \quad \text{stock} \\ \mu &= \text{constant for all } t. \end{aligned}$$

The labor series was obtained from an internal Bell System Study.¹³ The labor input is measured in terms of man-hours actually worked, adjusted for changes in labor quality. Labor quality is quantified by categorizing man-hours actually worked in the base period (1967) in terms of occupation definition and years of service. The relative weights obtained from the categorizing were held fixed and used to weight the man-hours of each subsequent year.

A production function was estimated using the capital and labor inputs as defined above and utilizing aggregate demand (assumed equal to output over the estimation interval) adjusted to a net value added (*NVA*) basis, i.e., the output produced by the primary factors of production.

Aggregate demand for service is a value-measured index which is equal to

$$Q_t = \sum_i Q_{it}$$

where

i = Toll, Local, Private Line, WATS and other Miscellaneous services.

¹⁰ In [11].

¹¹ The retirement series (needed to generate capital under the above schema) was constructed from Bell System data. Due to the assumptions above, i.e., the "one-horse shay" assumption on capital decay, a uniform distribution for equipment lifetimes was found with an average life of Bell System plant of 23 years.

¹² It is recognized that this assumption does not hold in the real world and in particular utilization is generally considered to be a managerial control variable to an extent. However, no aggregate proxy was found to remedy this situation, although considerable work in this area is on-going.

¹³ In [1].

Net Value Added is defined as:

$$NVA(t) = Q_t - \text{Material Costs} - \text{Nonrecurring Costs} - \text{Indirect Taxes} - \text{Depreciation Expenses.} \quad (16)$$

(All terms are in constant dollars.)

The specification of these intermediate expenses will be made in the Financial Module.

The production function found to yield the best performance was an aggregate homogeneous production function with neutral technological change using the Net Value Added (*NVA*) measure of output.¹⁴ The form is

$$NVA(t) = A[K(t)]^\alpha [L(t)]^\beta e^{\gamma t}. \quad (17)$$

The econometric estimate was found (in log linear form) to be

$$\begin{aligned} \log NVA = & -2.504 + 0.545 \log K(t) \\ & (1.35) \quad (0.155) \\ & + 0.549 \log L(t) + 0.053 \text{ Time} \quad (17a) \\ & (0.0776) \quad (0.005) \\ \bar{R}^2 = & 0.999 \quad D.W. = 1.61 \quad SER = 0.011. \end{aligned}$$

The period of fit is from 1954 to 1970, annually.

Statistically this equation performs quite well. The coefficients of capital, labor, and technology are all statistically significant and are at magnitudes consistent with other similar studies.¹⁵ The sum of α and β which is the measure of economies of scale is equal to 1.09, which indicates increasing returns to scale. Statistically, this sum is significantly different from unity at the .05 level of significance.

The mode of operation of the Production Module is that it receives the quantity demanded from the Demand Module as an input; from this labor and capital must be determined. There is therefore no way to choose labor and capital unambiguously, without imposing additional conditions. The factor mix process in the Input Module of the Management Model is used to supply these additional conditions.

Four options are provided:

- (1) Choose the desired level of capital. An example of this mode of operation would be to use an exogenous Construction Program as an input variable to specify the capital level. The labor level is then derived to meet demand.
- (2) Choose the desired level of labor. An example of this mode of operation would be to use an administrative directive of "no new hiring" as an input variable to specify the labor level. The amount of new construction (capital level) which allows demand to be satisfied can then be found.
- (3) Maintain the traditional capital-labor input ratio. In this mode, the capital level and labor level are simultaneously determined in a "Keynesian fashion" as demand-derived quantities.

¹⁴ More exotic forms of production functions were tried but none performed better than this most simple form. See Jorgenson [10] for a review of production function fitting.

¹⁵ See Dobell et al. [7] as a recent example.

- (4) Choose the capital-labor mix by a minimum-cost strategy. In this mode, comparisons between traditional capital-labor mixes and the minimum cost can be made.

Options (1) and (2) provide the user with a management procedure for investigating alternative methods of meeting demand, as well as with a mechanism for choosing output which is less than or greater than that demanded. In the latter case, lost revenues can be calculated which provide a measure of the penalty for not meeting or for over-meeting demand. Option (3) yields an equation of the historical capital-labor ratio which, when used simultaneously with the production function, allows the factor mix and level to be derived from demand. Option (4) is a surrogate for input optimization where factor prices—wage rate and cost of capital—enter into the determination of factor mix. As their respective prices change, the minimum-cost mix of labor and capital is defined by

$$\log [K/L]_t - \log [K/L]_{t-1} = E\{\log [\omega/\rho]_t - \log [\omega/\rho]_{t-1}\}, \quad (18)$$

where

ρ = Cost of capital,

ω = Wage rate, and

E = Elasticity of input substitution, a constant estimated from historical data (expected value = 1).

The Production and Financial Modules interact, since the method of choice in financing the level of capital K will affect its cost ρ .

Financial module

The principal function of the Financial Module (see Figure 4E) is to determine the total expenses associated with operation under a set of economic conditions derived from the National Economy Module, using a prescribed investment strategy derived from the Management Model. These total expenses are of two types: (1) indirect expenses and taxes needed to compute the Net Value Added as well as total operating expenses; and (2) financial "expenses" such as interest charges, etc. The former use econometric relationships; the latter are found by accounting identities.

The net value added formulation requires estimation of indirect expenses and taxes other than income tax. Net Value Added (NVA) is defined as

$$NVA(t) = Q(t) - \frac{APTAX(t)}{PGNP(t)} - \frac{AOTAX(t)}{PGNP(t)} - \frac{ADEP(t)}{PIFIXER(t)} - \frac{AMRAS(t)}{PGNP(t)} \quad (19)$$

where

$Q(t)$ = Constant dollar gross output,

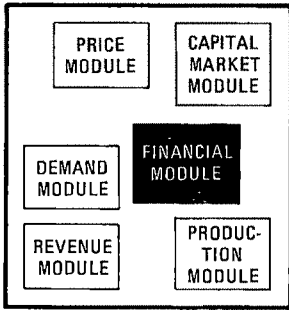
$APTAX$ = Property tax,

$AOTAX$ = Other taxes,

$ADEP$ = Depreciation expenses,

$AMRAS$ = Expenses for materials, rents, services,

FIGURE 4E
CORPORATE MODEL



$PGNP$ = Implicit price deflator for GNP, and

$PIFIXER$ = Implicit price deflator-private fixed nonresidential investment.

The following econometric equations are used for determining these operating expenses.

Materials, Rents, and Services—AMRAS

$$\log(AMRAS) = 12.8113 + 1.18963 \log[PXEAF(-1)] \\ (0.05014) (0.180627) + 0.06750 \text{ Time} \quad (20) \\ (0.003494)$$

$$\bar{R}^2 = 0.9984 \quad D.W. = 1.7708 \quad SER = 0.015268,$$

where

$PXEAF$ = Price of gross product — nonfarm industries.

The period of fit is from 1958 to 1971, annually.

Property Tax—APTAX

$$\log(APTAX) = -10.8029 + 1.76326 \log[ATPC(-1)] \\ (1.27872) (0.144145) + 0.054386 \text{ Time} \quad (21) \\ (0.011074)$$

$$\bar{R}^2 = 0.9999 \quad D.W. = 1.6972 \quad SER = 0.003279.$$

The period of fit is from 1960 to 1970, annually.

Property taxes are shown as a function of the one period lagged value of Average Telephone Plant ($ATPC$) and a time trend. Average Telephone Plant is estimated from the capital series as:

Average Telephone Plant—ATPC

$$ATPC = -24275.0 + 1.58286 K \\ (792.1) (0.0192) \quad (22)$$

$$\bar{R}^2 = 0.999 \quad D.W. = 1.88 \quad SER = 266.9.$$

The period of fit is from 1964 to 1971, annually.

Since $K(t)$, the capital series, is forecasted through the use of the production function, $ATPC$ is also forecasted. In a sense, this relationship simply translates the economic measure of capital stock to the accounting relationship Average Telephone Plant.

Other Taxes—AOTAX

$$AOTAX = 61.3633 + 3.2982 TGSL \\ (41.7264) (1.14282) + 0.387425 AOTAX(-1) - 5.66205 \text{ Time} \quad (23) \\ (0.25517) (3.5345)$$

$$\bar{R}^2 = 0.9966 \quad D.W. = 1.4998 \quad SER = 7.78,$$

where

$TGSL$ = State and local government tax receipts.

The period of fit is from 1960 to 1971, annually.

Depreciation Expenses—*ADEP*

Due to the observed empirical trend to an approximate average plant life, depreciation expenses are estimated by the simple relationship

$$ADEP = 0.55565 \text{ } ATPC(-1), \quad (24)$$

where

$ATPC(-1)$ is Average Telephone Plant lagged one year.

With the above, Net Value Added can be computed. Two additional expenses are needed to compute total operating expenses. These are labor expenses and Social Security taxes.

Labor expenses are found by multiplying the number of weighted man-hours required by the forecast implicit wage rate. This rate represents an aggregate Bell System cost per weighted man-hour of effort. The Wage Rate equation is given by:

Wage Rate (*RATE*)

$$\begin{aligned} \log(RATE) = & 0.703417 + 0.94096 \log(PXEAF) \\ & (0.308724) \quad (0.249602) \\ & + 0.35867 \log[RATE(-1)] + 0.34366 \log(NVA/L) \\ & (0.294339) \quad (0.022574) \\ & + 0.068731 \log[NVA(-1)/L(-1)] \\ & (0.04714) \\ & + 0.10310 \log[NVA(-2)/L(-2)] \\ & (0.06772) \end{aligned} \quad (25)$$

$$\bar{R}^2 = 0.9964 \quad D.W. = 2.2026 \quad SER = 0.0111.$$

The period of fit is from 1958 to 1970, annually.

The aggregate (implicit) Bell System wage rate (which includes payments for pensions, benefits, etc.) is estimated as a function of the inflation level given by *PXEAF* (Price of gross products-nonfarm industries) and labor productivity. The productivity measure is defined by output (*NVA*) per weighted man-hour. An examination of *RATE* shows that a unit change in the inflation level induces a 0.94 unit change in wages, whereas an aggregate unit change in productivity distributed over a three-year period results in a 0.137 unit change in the wage rate. The wage rate includes all labor costs (pensions, death benefits, etc.) except Social Security expenses. Social Security expenses are forecasted by the following:

Social Security Tax

$$SS - SS(-1) = 3.83384[TW - TW(-1)] \\ (0.24008) \quad (26)$$

$$\bar{R}^2 = 0.8482 \quad D.W. = 1.8187 \quad SER = 4.0112,$$

where

TW = Contributions for Social Security (a macro variable).

The period of fit is from 1948 to 1971, annually.

The principal remaining function of the Financial Module is the computation of interest expenses and traditional "book" relation-

ships for Income Statement purposes. Using accounting identities, in particular, those needed to convert marginal interest charges to "embedded" interest rates, book values of debt and equity, etc., each of these items is computed. In addition, mechanisms which allow alternative investment strategies are also provided. There are four investment strategy options within the Financial Module, each providing a variation in financial mix (see the Management Model in Figure 2). Retention, new equity, or new debt can be varied at the discretion of the user. No provision is currently available expressly for "convertible instruments." One final option is presently being developed which will choose the financial mix which maximizes earnings per share. This option works in simultaneous solution with the Capital Market Module so that retention, debt structure, and new flotations are simultaneously chosen.

Capital market module

The Capital Market Module (see Figure 4F) is a characterization of the capital markets and financial world, within which the corporation must operate. The central behavioral equations are attempts to characterize the process of market valuation of both debt and equity (hence, cost of capital).

One principal endogenous quantity determined by this module is $\rho(t)$, the total cost of capital. It is defined as the market weighted sum of marginal cost of equity, plus the marginal cost of debt:

$$\rho(t) = \frac{\gamma(t) \cdot S(t)}{S(t) + B(t)} + \frac{i(t) \cdot B(t)}{S(t) + B(t)}, \quad (27)$$

where

$\gamma(t)$ = Equity cost of capital,

$i(t)$ = Debt cost of capital,

$S(t)$ = Total market value of equity, and

$B(t)$ = Total market value of debt.

The market value of debt, $B(t)$, is defined as

$$B(t) = D(t - 1) \left[\frac{RD(t - 1)}{i(t)} \right] + I_D(t) - R_D(t), \quad (28)$$

where

$D(t)$ = Book value of debt at time t ,

$RD(t)$ = "Embedded" interest rate of book debt at t ,

$I_D(t)$ = New debt issue at time (t) , and

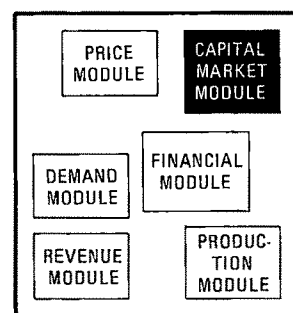
$R_D(t)$ = Debt retirements at time (t) .

The equity cost of capital is not so directly determined. It is estimated indirectly by finding the value of equity by using jointly a variant of the Sharpe-Lintner-Mossin capital asset pricing model¹⁶ and the Modigliani-Miller construct of equilibrium valuation.¹⁷

The value of equity by the Capital Asset Pricing Model is

$$\frac{S(t + 1) - S(t)}{S(t)} = \alpha + \beta I_1(t) + \delta I_2(t), \quad (29)$$

FIGURE 4F
CORPORATE MODEL



¹⁶ See [17].

¹⁷ See [13].

where

$S(t)$ = Total market value of A.T.&T. at time t ,

$I_1(t)$ = Percentage rate of change of a Capital Market Index # 1,
and

$I_2(t)$ = Percentage rate of change of a Capital Market Index # 2.

This equation has been estimated and is:

$$RMVAL = 0.3736RS\&P - 0.47988 RINT \quad (29a)$$

(0.12535) (0.16471)

$$\bar{R}^2 = 0.43652 \quad D.W. = 1.77 \quad SER = 0.0381,$$

where

$$RMVAL = \frac{S(t+1) - S(t)}{S(t)},$$

$RS\&P$ = Rate of change of the Standard and Poor's Index, and

$RINT$ = Rate of change of the market interest rate taken as the
Aaa rate.

The period of fit is from 1962:4 to 1971:4, quarterly.

This equation reflects the positive effect of the stock market on the market value of A.T.&T., and the converse impact of interest rate changes.

The above is a single period model under conditions of equilibrium in the capital markets. A consistent alternative representation of $S(t+1)$ is given below:¹⁸

$$S(t+1) = \frac{\bar{RE}(t+1)[K(t) - D(t)]}{\gamma(t+1)} + \frac{[I(t+1) - I_D(t+1)][\tilde{RE}(t+1) - \gamma(t+1)]}{\gamma(t+1)} \quad (30)$$

where

$\bar{RE}(t+1)$ = Expected rate of return on existing equity capital in
 $t+1$,

$\tilde{RE}(t+1)$ = Expected rate of return on new equity investment,

$\gamma(t+1)$ = Equity cost of capital, and

$I(t+1)$ = Total new investment in $t+1$.

To derive the equity cost of capital, the two equivalent representations of valuation are equilibrated. With this the over-all cost of capital can be derived.

For forecasting the cost of capital the same procedure as above is used. The two forecasted representations of valuation are equilibrated yielding $\gamma(t)$. The marginal cost of debt capital, $i(t)$, is forecasted and the two are combined as prescribed above to give the forecast of $\rho(t)$, the overall cost of capital. To augment this approach an alternative representation for $\gamma(t)$ is used to discern the impact of policy on equity cost of capital. In effect, the *ex post* series of equity capital

¹⁸ See Hamada [8] for a proof of their equivalency.

costs was used to derive an equation in terms of variables which significantly affect equity costs. The equation is:

$$\gamma(t) = \frac{0.00251}{(0.00151)} + \frac{0.25021i(t+1)}{(0.13176)} + \frac{0.005784\lambda(t)}{(0.00333)} + \frac{0.447863}{(0.17130)} \gamma(t-1) \quad (31)$$

$$\bar{R} = 0.8151 \quad D.W. = 2.2664 \quad SER = 0.00127,$$

where

$$\lambda(t) = \text{debt-equity ratio at time } t.$$

The period of fit is from 1965:1 to 1971:2, quarterly.

The equation indicates that the cost of equity capital increases as both the anticipated interest levels [$i(t+1)$] and the debt-equity ratio [$\lambda(t)$] increase, and is strongly influenced by the past cost. This relationship is used in policy simulations where the effects of different investment strategies are reviewed in terms of overall capital costs.

With the above, equilibrium share price is computed as

$$P(t+1) = \frac{S(t+1)}{N(t+1)}. \quad (32)$$

If an equity issue is used in the investment strategy, then it is currently assumed that the issue will be floated at the prevailing market price and that the market capitalizes this action. With this, an alternative form for $P(t+1)$ is¹⁹

$$P(t+1) = \frac{S(t+1) - I_s(t+1)}{N(t)}, \quad (33)$$

where

$$I_s(t+1) = \text{New Equity financed by the stock issue.}$$

The number of shares required to receive the amount $I_s(t+1)$ is

$$n(t+1) = \frac{I_s(t+1)N(t)}{S(t+1) - I_s(t+1)}. \quad (34)$$

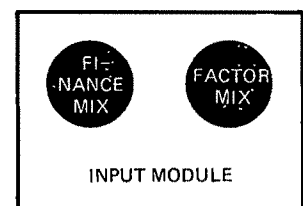
Thus the expected number of shares to be issued, $n(t+1)$, is given to the Financial Model so that earnings per share can be forecast under conditions of equity flotation [$N(t+1) = n(t+1) + N(t)$].

This completes the specification of the Corporate Model. The remaining Model within FORECYT is the Management Model.

□ **Management model.** The input of the Management Model translates the Corporate variables concerned with pricing, investment, and finance into policy. See Figure 5. These variables are at the control of the user, who can establish his own policy, so as to maintain historical relationships, or to test alternatives proposed by others.

The effectiveness of these alternative strategies for a given set of macroeconomic conditions is measured in terms of a set of overall quantitative indicators of Corporate performance, which are dis-

FIGURE 5
MANAGEMENT MODEL



¹⁹ See [13] for a proof of this relationship.

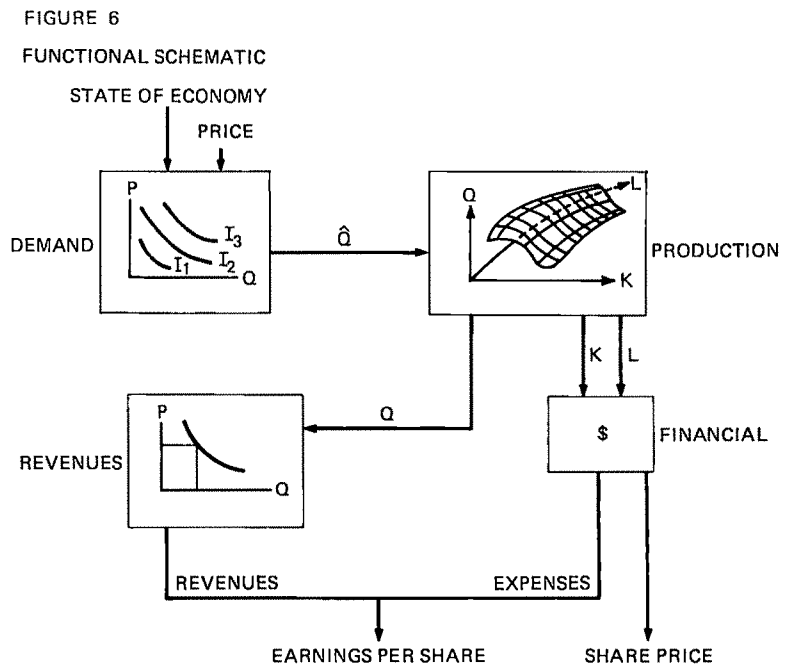
played in the Output Module. No attempt at a single "figure of merit" has been made; however, various combinations of the indicators can be constructed by the "decision maker."

The quantitative measures of policy are those which have been historically viewed as indicative of "well-being." Most are of the traditional balance sheet variety, such as total revenue, total expenses, net revenue, and earnings per share. Others are indicators of the difference between potential and realized performance, such as lost sales, or traditional business ratios (i.e., adjusted labor expense per dollar of adjusted revenue and adjusted dollar of productive capacity per dollar of adjusted revenue). Collectively, these indicators provide a basis for observation, appraisal, and pattern study of proposed Corporate policies.

4. Operation of the model

■ FORECYT can be utilized for analysis with a number of options. Figure 6 depicts the mechanism of operation for forecasting in the following manner:

- (1) Forecasts are generated by the National Economy Module to provide a number of primary inputs (shown as "the State of the Economy") to the Corporate Model:
 - (a) Forecasts of National Income Accounts are a primary driving element from which gross demand for Bell System services is estimated, and
 - (b) Forecasts of financial variables are used to prescribe capital market conditions and general financial conditions facing the corporation.
- (2) A prescribed regulatory posture shown as "Price" is generated by the Regulatory Module. From these inputs, demand can be derived in a forecasting sense. The derived demand, $\hat{Q}$, drives the Production Module which forecasts the levels of capital and labor.



- (3) Depending on the financial mix and factor (labor-capital) mix, investment and labor force augmentation determine the demand, Q , actually provided, from which revenues are determined.
- (4) The Financial Module, interacting with the Capital Market Module, determines anticipated total expenses and share price. At this stage, after the above steps, the measures of performance indicated in the Output Module, such as earnings per share, are then computed.

In the forecasting mode the normal interval of forecast is quarterly for eight quarters. To a large extent, this is predicated on the forecasting intervals of the national economy models used. For planning analysis studies, the interval can be longer. Using the same model and driving mechanisms, analyses over five- and ten-year periods are made—albeit, less precise and more speculative than those for eight quarters ahead. However, it is here that the long run consequences of proposed alternative courses of action can be evaluated. In this mode of operation the trade-off is between precision in the result and the ability to glimpse, crudely but logically consistently, the consequences of alternative futures.

■ A key operational mode of FORECYT is to establish a control solution from which policy analyses can be made. This requires a forecast using the “best guess” of how the national economy will evolve as well as specifications of “normal” anticipated corporate policy. To demonstrate the performance of FORECYT in this Control Solution mode the period of 1972 will be examined by using a forecast made in January, 1972 and by comparing it to actual Bell System performance of a number of key corporate variables over the same interval.

In January, 1972, a forecast of Bell System performance for 1972 was made using FORECYT. The forecast of the economy used was provided by Data Resources, Incorporated²⁰ and was their eight-quarter solution of the U. S. economy. In addition, regulatory posture was assumed to be such that there would be 5-percent annual increase in the local and message toll price indices with no changes in the other price indices. Also, it was assumed that there would be no common equity flotation and that A.T.&T. would pay the Aaa bond rate for debt flotations. Table 2 shows some results of this forecast compared to actual performance for the four quarters of 1972. The aggregate revenues, expense, and tax quantities shown in Table 2 result from forecasting the individual components using the system of equations detailed in Section 3. Net Income and Earnings per Share, however, are computed from these forecasted quantities. The error for these will in general be larger since they are essentially the difference between two large numbers of nearly the same magnitude. Overall, the forecasting precision appears acceptable.

It should be noted that there are two sources of error in the forecast which should not be attributed to the model. The macro variables are forecasts and the price changes are forecasts as well as the marginal interest rate. Table 3 is an “adjusted” forecast using

5. Evaluation of model performance

²⁰ In [4].

TABLE 2

COMPARISON OF JANUARY 1972 FORECAST TO ACTUAL PERFORMANCE

FORECAST ITEMS	72:1			72:2			72:3			72:4		
	ACTUAL	ESTIMATE	%DIFF.	ACTUAL	ESTIMATE	%DIFF.	ACTUAL	ESTIMATE	%DIFF.	ACTUAL	ESTIMATE	%DIFF.
TOTAL OPERATING REVENUES & OTHER INCOME*	5135	5091	-0.86	5253	5218	-0.66	5437	5356	-1.49	5593	5491	-1.85
TOTAL OPERATING EXPENSES*	3254	3213	-1.20	3336	3293	-1.30	3419	3380	-1.14	3513	3465	-1.37
TOTAL OPERATING TAXES*	923	898	-2.70	923	919	-0.43	968	940	-2.90	981	959	-2.24
INTEREST EXPENSES*	350	350	+0.05	369	361	-2.16	380	380	-0.10	396	398	+0.50
NET INCOME*	609	630	+3.40	624	644	+3.20	669	657	-1.80	702	668	-4.80
EARNINGS PER SHARE	1.019	1.045	+2.50	1.046	1.070	+2.30	1.108	1.093	-1.35	1.168	1.113	-4.70
*FIGURES IN MILLIONS OF DOLLARS												

FORECYT where actual U. S. economy performance, actual price changes, and interest charges were used. The performance here is better and is indicative of a significant source of potential error in planning analysis. It does not suffice to predict the internal posture of a corporation when planning. The environment critically affects anticipated outcomes. All forecasts are predicated on an assumed

TABLE 3

COMPARISON OF "ADJUSTED" FORECAST TO ACTUAL PERFORMANCE

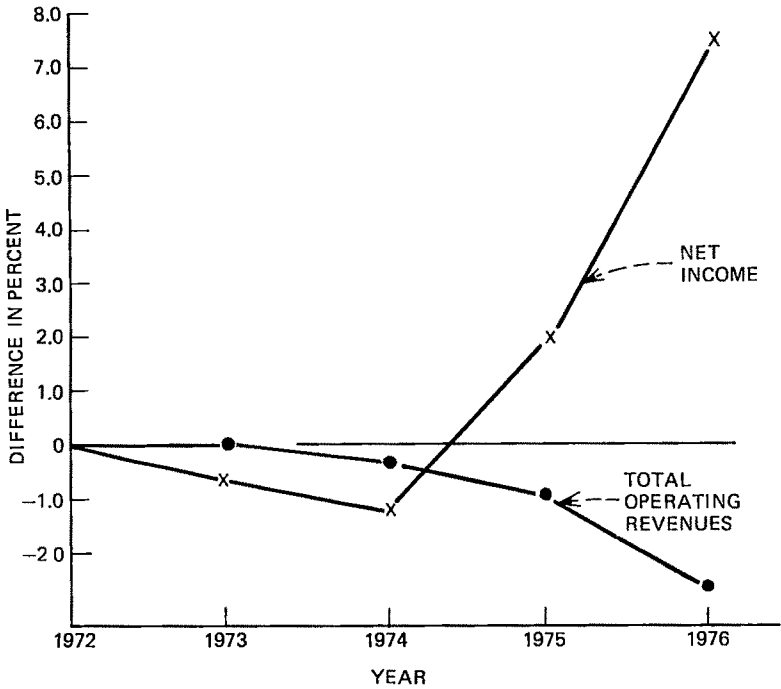
FORECAST ITEMS	72:1			72:2			72:3			72:4		
	ACTUAL	ESTIMATE	%DIFF.	ACTUAL	ESTIMATE	%DIFF.	ACTUAL	ESTIMATE	%DIFF.	ACTUAL	ESTIMATE	%DIFF.
TOTAL OPERATING REVENUES & OTHER INCOME*	5135	5097	-0.74	5253	5240	-0.25	5437	5376	-1.12	5593	5531	-1.11
TOTAL OPERATING EXPENSES*	3254	3208	-1.40	3336	3302	-1.01	3419	3359	-1.70	3513	3429	-2.39
TOTAL OPERATING TAXES*	923	912	-1.12	923	932	+0.97	968	968	+0.05	981	1005	+2.45
INTEREST EXPENSES*	350	350	-	369	362	-1.95	380	381	+0.20	396	400	+1.00
NET INCOME*	609	627	+2.95	624	643	+3.04	669	669	-0.09	702	697	-0.71
EARNINGS PER SHARE	1.019	1.039	+1.96	1.046	1.068	+2.10	1.108	1.109	+0.09	1.168	1.159	-0.77
*FIGURES IN MILLIONS OF DOLLARS												

B. E. DAVIS

G. J. CACCAPPOLO AND

50 / M. A. CHAUDRY

FIGURE 7
COMPARISON BETWEEN ECONOMY 1 AND ECONOMY 2



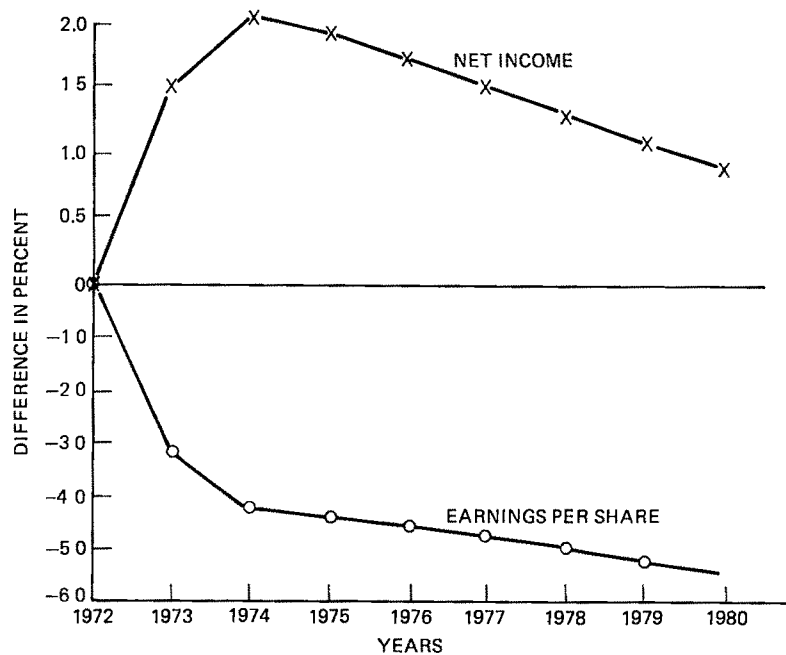
future state of the world. Econometrics only allows these assumptions to be quantified.

To demonstrate the possible effect of different states of the U. S. economy a long-run analysis will be examined where the corporation is faced with the possibility of two quite different future “states of the world.” The details of the economy are not important for this exposition except in that Economy 1 is characterized by stimulated demand and its accompanying higher inflation while Economy 2 is one of restraint with controlled inflation.²¹ The corporate policy variables in each case were as indicated for the 1972 forecast. The analysis was run from 1972 to 1976 using adjustments to the long-term macroeconomic forecasts of the U. S. economy developed by Data Resources, Incorporated.²² Figure 7 shows partial results of this analysis. The percent differences are defined with respect to Economy 2; thus a negative percent difference in total operating revenues indicates that operating revenues are forecasted to be high in Economy 1 with respect to Economy 2. As one would expect, the total Operating Revenues under the stimulated economy (Economy 1) always exceed the revenues under Economy 2. Perhaps the most interesting result from this scenario is the switch in Net Income performance under the different economies. Initially (1973 and 1974), Net Income under Economy 1 exceeds that from Economy 2. However, from 1974 on, inflation and its attendant increased expenses take their toll for Economy 1. The result is, in a sense, intuitive since it is widely believed that inflation is not kind to public utilities (or other capital intensive industries, in fact). This analysis simply affirms this belief with an interesting “start-up” phenomenon.

As a last demonstration of a possible planning scenario a single

²¹ Both economies were developed as simulations by Data Resources, Incorporated.
²² In [5].

FIGURE 8
COMPARISON BETWEEN INVESTMENT STRATEGIES



equity issue was simulated for Bell System financing for second quarter 1973. The issue price was set arbitrarily slightly below book-value and the number of shares issued was approximately 6 percent of the number of shares outstanding. The impact of this issue is compared to a solution where no stock issues are made. All other variables are held the same. Figure 8 shows some partial results of this analysis over the interval of 1973 to 1980. They are defined with respect to the issue case. Hence, a negative percent difference in earnings per share indicates earnings per share are lower for the issue case.

This example as constructed is, in essence, a shock analysis since there is a single issue and then never another one. The results show two important and predictable consequences. First, the advantage of a stock issue below book value is unfavorable to the individual share owner while favorable to the firm as a whole (Earnings per Share and Net Income, respectively). Notice also that the gains from the issue to the firm are eroding over time as the advantage in capital expenses (deferred interest charges) is disappearing.²³ Notice, however, that the dilution effect on earnings per share is persisting and it will approach a level of 6.33 percent as time becomes very large. Thus, as theory would predict, the effects of dilution do not disappear but persist, all other things equal. Additional analysis could indicate which issue price would not dilute the shareowner's position.

6. Summary

B E DAVIS
G. J. CACCAPPOLO AND
52 / M. A. CHAUDRY

■ An econometric model for corporate planning analysis has been discussed and demonstrated. The results of the policy analyses were what theory would have predicted. The structure and econometrics simply allow the theory to be examined quantitatively. In a sense,

²³ For very large t the difference will approach zero.

this has been the theme of this paper; that is, for present day corporate planning modeling, the state of applied economics and the availability of planning technology is such that comprehensive and informative models can be constructed on simple economic theory.²⁴

Appendix

■ **PLOC-local service price index.** *PLOC* is the Local Telephone Service index²⁵ calculated for the System for both residence and business, and is divided into four major price categories: main stations, extensions, additional message units, and service connections (moves and changes.) In addition, the index is disaggregated by state. This results in the definition of 400 separate commodity groups. The index is then computed to be

$$PLOC_{s(t)} = \frac{\sum_{i \in I} q_{i0} P_{it}}{\sum_{i \in I} q_{i0} P_{i0}}, \quad (A1)$$

where

q_{i0} = Quantity sold of the i th service in the base period,

P_{it} = Price of the i th service in the time interval t , and

$PLOC_{s(t)}$ = Local price index for state s at time t .

Since detailed subdivisions are not generally available, groups of services were constructed such that a representative service could reflect the price movement of that group. In particular, if a group of services J has prices which change uniformly proportionally, then for all K services in group J

$$\sum_{i \in J} q_{i0} P_{it} = \frac{P_{Kt}}{P_{K0}} \left(\sum_{i \in J} q_{i0} P_{i0} \right). \quad (A2)$$

Thus

$$PLOC_{s(t)} = \frac{\sum_{k \in I} \frac{P_{Kt}}{P_{K0}} \left(\sum_{i \in J} q_{i0} P_{i0} \right)}{\sum_{i \in I} q_{i0} P_{i0}}.$$

Hence, the contribution of the service group can be estimated using only the base revenues for the group and the price series for the representative service within that group.

The System's Local Telephone price index was computed as a weighted sum of each of the local state indices:

$$PLOC_t = \sum_s \lambda_s \cdot PLOC_{s(t)}, \quad (A3)$$

where

$$\sum \lambda_s = 1, \text{ and}$$

$$\lambda_s = \text{Proportion of total local revenues from state } s.$$

Description of price indices

²⁴ See [14] and [12] for additional discussion on this topic.

²⁵ The development of the local service price index, *PLOC*, was done by R. L. Shanahan [15] who has analyzed this service offering in a more comprehensive nature than presented here.

□ **PTOL-message toll price index.** The Message Toll price index²⁶ is a weighted sum of

$PINTER$ = Interstate Message Toll Service price index, and

$PINTRA$ = Intrastate Message Toll Service price index.

The formulas are:

(1) Interstate Message Toll [$PINTER$]

$$PINTER_t = \frac{\sum_{i=1}^5 \sum_{j=1}^4 \sum_{k=1}^3 W_{ijk(1970)} \cdot P_{ijk(t)}}{\sum_{i=1}^5 \sum_{j=1}^4 \sum_{k=1}^3 W_{ijk(1970)} \cdot P_{ijk(1970)}}, \quad (A4)$$

where

i = Mileage bands (1, 2, . . . , 5),

j = Time of day (1, 2, 3, 4),

k = Type of call (1, 2, 3),

W_{ijk} = Volume of interstate calls for the (ijk) th category for 1970:4, and

$P_{ijk(t)}$ = Price for a three-minute call of the (ijk) th category for period t .

(2) Intrastate Message Toll [$PINTRA$]

Individual State Price Index

$$PINTRA_{s(t)} = \frac{\sum_{i=1}^5 \sum_{j=1}^4 \sum_{k=1}^3 V_{sijk(1969)} \cdot P_{sijk(t)}}{\sum_{i=1}^5 \sum_{j=1}^4 \sum_{k=1}^3 V_{sijk(1969)} \cdot P_{sijk(1969)}}, \quad (A5)$$

where

s = State (1, 2, 3, . . .),

i = Mileage band (1, 2, . . . , 5),

j = Time of day (1, 2, 3, 4),

k = Type of call (1, 2, 3),

$P_{sijk(t)}$ = Price of intrastate toll call of (ijk) th category made within state s in period t , and

V_{sijk} = Volume of intrastate toll calls of (ijk) th category in state s in the base period, (1969:4).

The Bell System Intrastate Toll Price Index is then defined as:

$$PINTRA_t = \sum \lambda_s \cdot PINTRA_{s(t)} \quad (A6)$$

$$\sum \lambda_s = 1,$$

where

λ_s = Proportion of the total System intrastate calls made within state s .

The overall Message Toll Service index is the weighted sum of $PINTER$ and $PINTRA$:

$$PTOL_{(t)} = \alpha \cdot PINTER + (1 - \alpha) \cdot PINTRA, \quad (A7)$$

where

α = Proportion of the toll revenues due to interstate demand during the base period.

□ **PWATS-WATS price index.** The WATS price index is a weighted average of prices for full-time and measured-timed outward WATS (interstate) lines in 58 WATS states:

$$PWATS = \{ \lambda_f [\sum_{i=1}^6 \sum_{s=1}^{58} l_{ifs} \cdot P_{ifs(t)}] + (1 - \lambda_f) [\sum_{i=1}^6 \sum_{s=1}^{58} l_{ims} \cdot P_{ims(t)}] \} / \quad (A8)$$

$$\{ \lambda_f [\sum_{i=1}^6 \sum_{s=1}^{58} l_{ifs} \cdot P_{ifs(1970)}] + (1 - \lambda_f) [\sum_{i=1}^6 \sum_{s=1}^{58} l_{ims} \cdot P_{ims(1970)}] \},$$

where

i = WATS zones (1, 2, . . . , 6),

s = WATS state (1, 2, . . . , 58),

P_{ifs} = Price of a full-time line for i th zone in state s ,

l_{ifs} = Number of full-time lines of i th zone in state s in 1970,

P_{ims} = Price of a measured-time line of zone i in state s ,

l_{ims} = Number of measured-time lines of zone i in state s ,

λ_f = Proportion of WATS revenues derived from full-time WATS lines in 1970, and

$1 - \lambda_f$ = Proportion of WATS revenues derived from measured time lines (including overtime) in 1970.

□ **PPL-private line price index.** The Private Line price index²⁷ is a weighted average of Private Line and TELPAK rates in 91 selected point-to-point markets:

$$PPL = [\sum_{i=1}^{91} W_i P_{i(t)} / \sum_{i=1}^{91} W_i \cdot P_{i(1970)}] \times 100, \quad (A9)$$

where

W_i = Number of Private Lines in market i in 1970, and

P_{it} = Weighted price of a Private Line and a TELPAK line in year t (weights being the proportion of circuit miles in each of the types of line in 1970).

²⁷ See Chaudry [2].

References

1. AMERICAN TELEPHONE AND TELEGRAPH COMPANY. *Bell System Productivity Study—1947–70*. New York: American Telephone and Telegraph Company, October 1971.
2. CHAUDRY, M. A. "The Economics of Competition in the U. S. Communications Industry—An Econometric Approach." Unpublished Ph. D. dissertation. Tufts University, 1971.
3. ———. "Inter- and Intrastate Toll Price Indices for the Bell System: 1962–1971." Unpublished memorandum, American Telephone and Telegraph Company, April 1971.
4. DATA RESOURCES, INCORPORATED. *The Data Resources Review*. (December 20, 1971).
5. ———. "Ten-Year Projections of the U. S. Economy." *The Data Resources Review*, No. 7 (July 14, 1972).

6. DAVIS, B. E., CACCAPPOLO, G. J. AND CHAUDRY, M. A. "FORECYT: An Econometric Forecasting Model for Corporate Planning Analysis." *Corporate Planning Organization Report*. New York: American Telephone and Telegraph Company, June 1972.
7. DOBELL, A. R., TAYLOR, L. D., WAVERMAN, L., LIU, T. H., AND COPELAND, M. D. G. "Telephone Communications in Canada: Demand, Production, and Investment Decision." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 1 (Spring 1972), pp. 175-219.
8. HAMADA, R. S. "Portfolio Analysis, Market Equilibrium and Corporation Finance." *Journal of Finance*, Vol. 24, No. 1 (March 1969), pp. 13-31.
9. HOUTHAKKER, H. S. AND TAYLOR, L. D. *Consumer Demand in the United States*. 2nd ed. Cambridge: Harvard University Press, 1970.
10. JORGENSON, D. W. "Investment Behavior and the Production Function." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 1 (Spring 1972).
11. KUH, E. *Capital Stock Growth: A Micro-Economic Approach*. New York: North-Holland Publishing Co., 1963.
12. LEONTIEF, W. "Theoretical Assumptions and Nonobserved Facts." Presidential Address to the American Economic Association, December 29, 1970. Published in *The American Economic Review*, Vol. 61, No. 1 (March 1971).
13. MODIGLIANI, F. AND MILLER, M. H. "Dividend Policy, Growth, and the Valuation of Shares." *Journal of Business*, Vol. 34, No. 4 (October 1961), pp. 411-433.
14. PHELPS BROWN, E. H. "The Underdevelopment of Economics." Presidential Address to the Royal Economic Society, July 8, 1971. Published in the *Economic Journal*, Vol. 82, No. 324 (March 1972).
15. SHANAHAN, R. L. "Elasticity Effects in Lagged Demand Models." Unpublished memorandum, American Telephone and Telegraph Company, October 1972.
16. ———. "Local Exchange Price Index for the Bell System, 1961-1971." *Management Sciences Division Report*, American Telephone and Telegraph Company, April 1972.
17. SHARPE, W. F. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance*, Vol. 19, No. 3 (September 1964), pp. 425-442.
18. WHARTON SCHOOL, Econometric Research Unit. *Econometric Forecasting Model*. Philadelphia: University of Pennsylvania.

The behavior of a firm subject to stochastic regulatory review

Alvin K. Klevorick

Associate Professor
Cowles Foundation
Yale University

In recent years, the behavior of regulated public utilities has received increasing attention from economic theorists. The principal structure used in analyzing this behavior has been the Averch-Johnson model of the regulated firm. The current paper presents an alternative view of the regulated firm, one which attempts to narrow some (though by no means all) of the gaps between the now-standard model and the actual regulatory process. In particular, the new model proposed here considers the firm's operations in a dynamic context—with the firm looking to the future in making today's decisions—and it incorporates the interplay between the regulatory agency and the firm. The model captures the price-setting role of the regulators, and it encompasses the phenomenon of regulatory lag. The uncertainty associated with the occurrence of rate reviews is modeled by positing that reviews occur stochastically through time. And, although the treatment of the issue is rather simplistic, the model does incorporate technical change generated by the regulated firm's program of research and development. The regulated firm's optimal policy is characterized, and the implications this optimal policy has for two traditional issues in regulatory economics—the input efficiency of regulated firms and the effect of regulatory lag on research and development—are examined.

■ In recent years, the behavior of regulated public utilities has received increasing attention from economic theorists. The principal structure used in analyzing this behavior is the model of the regulated firm developed by H. Averch and L. L. Johnson (A-J) in 1962.¹ This now-standard model presents a static view of a profit-maximizing monopolist who faces a given demand curve for the single product

1. Introduction: beyond the Averch- Johnson model

Alvin K. Klevorick holds the B.A. degree from Amherst College (1963) and the M.A. and Ph.D. degrees from Princeton University (1965 and 1967, respectively). Dr. Klevorick has been with Yale University since 1967. His research interests are in economic theory and in law and economics.

The research described in this paper was carried out under grants from the National Science Foundation and the Ford Foundation. An Air Force Office of Scientific Research—National Research Council Postdoctoral Fellowship provided support for the author's work during 1969–70 when the basic ideas underlying the model in this paper were being formulated. His past discussions and joint research with William Baumol on the issues of public utility regulation have substantially shaped his current thinking on these matters. He is also grateful to Kim Peck, Herbert Scarf, Matthew Sobel, and Joseph Stiglitz for their helpful discussions concerning the present paper, and to Paul MacAvoy and Michael Rothschild for their helpful comments on an earlier version. These people, of course, bear no responsibility for any faults that may remain.

¹ In [1].

he produces, perfect markets for the two inputs (labor and capital) he uses in producing that product, and an explicit constraint on the rate of return he can earn. This last element of the regulated firm's environment—the "fair-rate-of-return" constraint—requires that the firm's net revenue (its gross revenue minus its operating costs) should not exceed a fixed percentage (the fair rate of return) of the value of the firm's capital stock net of depreciation (the rate base). The fair rate of return is bounded above by the return the firm would earn if it were able to maximize profits unconstrained, and it is bounded below by the market cost of capital.

The theoretical papers based on the A-J model² have provided a rather complete and detailed discussion of the input levels chosen by a regulated firm and the quantity of output produced by such a firm. For example, this literature has demonstrated that the profit-maximizing regulated firm will use a capital-labor ratio exceeding the one that minimizes cost for the output produced, and it has provided explicit comparisons of the output of the regulated monopolist with that of his unregulated counterpart. Moreover, several papers in this tradition have used the A-J model to determine (theoretically) how regulators might set the fair rate of return in order to induce regulated firms to behave in a social-welfare-maximizing way.

While the A-J model has been rich in analytical results, it is difficult to escape the fact that the model's picture of the regulatory process is quite rudimentary. Naturally, any model that hopes to yield analytical insights must abstract from the complicated reality of regulation and the regulated firm. But it is fair to ask at what price the A-J model's results have come. In compressing reality into a tractable model, have important aspects of the regulatory process been suppressed and, hence, their theoretical and policy implications been lost? The answer appears to be in the affirmative.

First, in the A-J model, the price-setting function of regulatory commissions, which is central to the working of regulation, has been tucked away and lost in the rate-of-return constraint. In reality, and contrary to the A-J model's depiction, the regulated firm does not simply face an ordinary demand curve along which it must choose a price-output combination subject to the fair-return constraint. Instead, the regulators tell the firm what price it can charge, and this price is presumably determined on the basis of some fair-rate-of-return calculation. As a public utility, the firm must then meet demand at the set price. The regulatory process, not the direct action of the market, adjusts this price upward or downward according to whether the firm is earning less or more than the fair rate of return. Hence, it is by the price that it must charge that the regulated firm is actually constrained and not, as the A-J model suggests, by some explicit fair-return constraint.

Second, the A-J model's completely static view of the regulated firm and the regulatory process also limits the model's ability to encompass some important regulatory issues. Once the firm in the model chooses its optimal position, specifically, its inputs of capital and labor (for then output and price follow from the production and

² Many of the papers based upon the Averch-Johnson model of the regulated firm are surveyed in Baumol and Klevorick [4]. Other papers that have used this model in discussing public utility regulation include Klevorick [5, 6], Sheshinski [7], and Wichers [8].

demand relationships), it remains at that position forever. In making this choice of inputs, the firm is assumed to treat capital and labor symmetrically, namely, as perfectly variable factors whose levels are to be chosen once and for all. There is no scope for growth or depreciation of the capital stock over time. One might be willing to assume away depreciation as an unnecessary complication in a model concerned primarily with other aspects of regulation. It is, however, more difficult to accept the symmetric treatment of labor and capital and the exclusion of capital expansion from a model explicitly concerned with the input efficiency of regulated firms.

The standard model's static vision stems, in part, from its assumptions that demand and production conditions are fixed and that input prices do not vary. On the demand side, there is no room for growth in demand. At the same time, the model ignores the possibility that the regulated firm might undertake research and development in an attempt to improve its production technology. Since technological progress would affect the terms on which labor and capital—the two inputs considered in the A-J framework—combine, research and development could have an important effect on the regulated firm's input efficiency. Attention to the role of technological change would require us to distinguish more carefully between the nearly perfectly variable factor, labor, and the more durable factor, capital, with further implications for the efficiency of the regulated firm's use of resources.

More fundamentally, though, the behavior of the firm in the A-J model lacks any dynamic character because the model assumes that the fair-rate-of-return constraint obtains as an equality at every point in time. The model states that if the fair rate of return is s percent, then the firm's net revenues are always exactly s percent of its rate base. This is incompatible with the observed fact that during some periods some regulated firms apparently earn more than what has been decided as the fair return, thereby giving rise to pressure from regulators for these firms to lower the prices charged for their services, while other regulated firms earn less than the fair rate, giving rise to permission from regulators for these firms to raise their prices.

Related to its assumption of a fair-rate constraint that is always effective is the A-J model's assumption that there is no regulatory lag. The model ignores the fact that the decisions of regulatory agencies do not change from one instant to the next but are, instead, revised only from time to time. Once the regulators set the price for the firm's output, the company must meet demand at that price until consumers or the firm request a new rate review and the regulators, if they see fit, reset the price.

Two implications of this intermittent nature of regulation bear emphasis. First, sometimes for substantial periods of time—the “regulatory lag” period—a regulated firm may be producing to meet demand at an “outmoded” price. That price was thought to yield a “fair return” under the demand and cost conditions prevailing when it was set but it is now “outmoded” because demand or technology or factor costs have changed. Second, the regulated firm does not know with certainty when it will be reviewed again. To be sure, if it is faring badly—say, because of unforeseen increases in factor costs—the firm can petition for a rate increase. But the firm cannot be certain how long any “excess profits” it is earning will last because it does not

know when some consumer group or the commission staff will initiate procedures for a rate reduction. The entire issue of the nature of regulatory lag and the beneficial or detrimental effect it may have on the input efficiency of regulated firms has been left outside the analytical framework of the A-J model. Those papers that have analyzed aspects of regulatory lag have found it necessary to step outside the A-J structure to do so.³

A final shortcoming of the standard model of the regulated firm is its assumption that perfect information is possessed by all participants in the regulatory process. The regulators in the A-J model know all there is to know about the firm's production and demand relationships while the firm knows exactly when it will be reviewed (namely, always) and exactly how the regulators will constrain it (namely, to earn exactly the fair rate of return). The information firms actually have about regulators and the information regulators have about the firms they control fall far short of the degree of perfection assumed in the A-J model.

The objective of this paper is to present a new model of the regulated firm, analyze the behavior of this firm, and reconsider some issues in regulatory economics in terms of this model. The new model attempts to narrow some of the gaps in the now-standard model of the regulated firm. In particular, the new structure proposed here considers the firm's operations in a dynamic context—with the firm looking to the future in making today's decisions—and it incorporates the interplay between the regulatory agency and the firm. The model captures the price-setting role of the regulators, and it encompasses the phenomenon of regulatory lag. The uncertainty associated with the occurrence of rate reviews is modeled by positing that reviews occur stochastically through time. And, although the treatment of the issue is rather simplistic, the model does incorporate technical change generated by the regulated firm's program of research and development. Clearly, the model developed here remains a gross oversimplification of the regulatory process as we know it. It is hoped, however, that by improving somewhat on the traditional basis for analysis, the new structure will provide new insights into the issues of regulation.

Section 2 presents the basic elements of the model and the mathematical statement of the regulated firm's problem in this model. The firm's optimal policy is characterized in Section 3, although the proofs of the theorems are left to the Appendix. Then Section 4 presents some implications of the model concerning the input efficiency of regulated firms and the role regulatory lag may be able to play in encouraging such firms to improve their technologies. Section 5 contains some concluding comments about directions for further work with this model.

2. A new model of the regulated firm

■ **Production and demand aspects of the firm's environment.** As in the A-J model, the regulated firm is assumed to be a monopolist producing a single homogeneous product. The company's output of this product in period t (q_t) is a function of the firm's capital stock

³ See, for example, Bailey [2], Bailey and Coleman [3], and Baumol and Klevorick [4].

in that period K_t , the labor it hires for the period L_t , and its stock of knowledge or its technology in that period A_t . While labor is a perfectly variable input, so that the firm can increase or decrease its labor input from one period to the next, capital (accumulated investment) and knowledge (accumulated research) are not perfectly variable—they can be increased but not decreased over time. In period t the firm chooses not only a labor input L_t but also an investment level I_t as an addition to its capital stock and an amount of research R_t by which it increments its stock of knowledge. There are, however, lags involved in putting new investment and technical knowledge to productive use. For simplicity it is assumed that investment and research in one period only become effective in the production process in the following period. Under this assumption of a one-period implementation lag and the assumption that capital does not depreciate and knowledge does not decay over time, the firm's productive capital stock in period $t + 1$ is simply $K_t + I_t$ while its stock of knowledge in that period is $A_t + R_t$.⁴

While the firm's ability to combine labor and capital at any point in time depends on the accumulated stock of knowledge, it is assumed that the firm's production function as a function of capital, labor, and the stock of knowledge is stable through time. This function is assumed to satisfy the following conditions:

$$q = F(K, L, A), \quad K \geq 0, \quad L \geq 0, \quad A \geq 0,$$

$$\frac{\partial q}{\partial K} > 0, \quad \frac{\partial q}{\partial L} > 0, \quad \frac{\partial q}{\partial A} > 0,$$

$$F(0, L, A) = F(K, 0, A) = F(K, L, 0) = 0.$$

It will be more convenient to express the firm's production capabilities in terms of its labor-requirements function, which is assumed to be twice continuously differentiable:

$$L = \mathcal{L}(q, K, A) \quad \text{with} \quad \frac{\partial \mathcal{L}}{\partial q} > 0, \quad \frac{\partial \mathcal{L}}{\partial K} < 0, \quad \frac{\partial \mathcal{L}}{\partial A} < 0. \quad (1)$$

The $\mathcal{L}(q, K, A)$ function indicates the amount of labor required to produce the output q with capital stock K and technology A , and it is obtained by solving the production function for L . The signs of the first partial derivatives of the labor-requirements function follow from the assumption that the marginal products of K , L , and A are always positive. The assumption that the firm's production relationships are stable through time means, essentially, that there is no exogenous technical change—neither the F function nor the $\mathcal{L}$ function depends on t as an independent variable. All changes in the firm's ability to combine capital and labor to produce output are the direct result of its own research undertakings.

In hiring labor and in undertaking investment and research, the firm is assumed to face perfect markets. Specifically, it costs the firm w monetary units to hire a worker for one period, c units of money to increase its capital stock by a unit in one period, and one

⁴ It would be easy to incorporate into the model a fixed depreciation rate for capital and a fixed decay rate for knowledge. To do so would, however, simply complicate the exposition without yielding any compensating analytical insights.

unit of money to do one unit of research in one period. At these prices per unit, which are assumed to remain constant through time, the firm can hire as many workers as it wants and undertake as much investment and research as it desires to pursue. The crucial distinction between the input markets in the current model and those in a natural multiperiod extension of the A-J model is that once the firm in the current model has added to its capital stock or its stock of knowledge, that plant and equipment and that knowledge remain with the firm forever. There is no market in which the firm can sell, and thereby reduce, its capital stock or its stock of knowledge. The input market is perfect with regard to purchases but not with regard to sales of stock inputs.

On the product side of the market, it is assumed that the price the firm receives for its product is related to its output level by an inverse demand function that remains fixed through time:

$$p = p(q), \quad p'(q) < 0. \quad (2)$$

Hence, the gross revenue the firm derives from the production and sale of output q is

$$G(q) = p(q)q, \quad (3)$$

and this function, which is assumed to be twice continuously differentiable, is also stable through time. The assumption of a static demand curve is rather restrictive. Especially in the context of public utility services, one would expect some growth in demand through time. The assumption can be removed but, on the whole, the additional insights gained at the level of analysis undertaken here do not justify the notational and expositional complications engendered.⁵ Hence, for the analysis in this paper, the demand function is assumed to be stable through time.

□ **The regulatory environment.** Having described the production and demand features of the firm's environment, it remains to describe how regulation proceeds in this model. The basic starting point is that regulators set the price the firm can charge for its service, and the responsibility of the regulated firm is to meet all demand at this price. Once the price is fixed by the regulators, it cannot be raised and hence—by virtue of a static demand curve—output cannot be lowered until the next regulatory review takes place. Consider the position of the firm at the start of period t . In the most recent review proceedings, the regulators set the firm's price at p_t . The firm is about to produce the output q_t demanded at this price using its capital stock K_t , its technology A_t , and the labor it has hired for the period L_t . Because of the capital and technology implementation lags, at the same time as it hires workers to produce for period t , the firm must choose its t th period investment and research levels. These will, in turn, determine the capital and technology it will have available in period $t + 1$.

At the start of period t , when it chooses L_t , I_t , and R_t , the firm does not know whether its position will be reviewed at the start of

⁵ There will, however, arise one point in the discussion of the firm's optimal policy when it would be more convenient to have demand growing, say at some constant rate or by some constant amount per period. At that point, the implications of such a growth assumption will be indicated and used to support a particular assumption about the optimal policy. See Section 4.

period $t + 1$. In reality, the probability that a firm is reviewed depends on the firm's recent history. If the firm has been making an inadequate return on its capital, it will probably petition for a price increase. On the other hand, consumer groups or the regulatory staff may call for a review on the grounds that the firm is earning "excess profits" and should be forced to lower its price. It is also probably true that a review is more likely in period $t + 1$ if the firm was not reviewed in period t . All these considerations are important. Nevertheless, as a first approximation, to gain some insight into the effect of the uncertain timing of reviews on regulated firm behavior, it will be assumed that regulatory reviews are generated by a stationary Bernoulli process. Specifically, the probability of a review—call it α —is assumed to be the same for any year, being independent of calendar time and the past history of reviews. The waiting time between reviews is then described by a geometric distribution, and a straightforward measure of regulatory lag is at hand—the expected value of the time between regulatory reviews: $1/\alpha$.

Hence, in planning its investment and research for period t , the firm believes that with probability α it will be reviewed at the start of period $t + 1$ and with probability $1 - \alpha$ it will not be reviewed at that time. In the absence of a review, the output price for period $t + 1$ will remain fixed at its period t level p_t . With the constant input prices and static demand conditions assumed in the model, an unchanged output price for period $t + 1$ means, for example, that if the firm did any cost-saving research in period t , it will have the temporary benefit of an excess earnings flow in period $t + 1$.⁶

While the firm is certain about what will ensue if it is not reviewed at the start of period $t + 1$, it cannot be sure what the outcome will be if it is reviewed at that time. The firm knows that if the regulators review its position, their goal will be to set a price that—given the firm's current demand and cost conditions—will yield the firm a "fair" rate of return on its current *capital* stock. If the regulators had perfect knowledge of the firm and its operations, it would be possible to be precise about their behavior. But the regulators do not know the firm's production technology or demand conditions perfectly. They cannot be sure precisely what effect a particular capital stock-technology pair will have on the firm's labor costs for different output levels. In addition, the regulators cannot be certain about the demand conditions, for example, the elasticity of demand, the firm faces.⁷ Hence, at best, the regulators take the information

⁶ This last point will be recognized, of course, as the heart of the argument of those who believe regulatory lag stimulates research and innovation. See, for example, Bailey [2], and Baumol and Klevorick [4], pp. 184–187.

⁷ The firm is also uncertain about aspects of its technology and the demand conditions it faces. For example, the firm cannot be sure what effect a particular level of research will have on its "stock of knowledge" and then on its productive capabilities. But the firm probably knows its demand and production conditions better than the regulators know them. It is assumed in this paper that the firm makes a point estimate of its demand function and its labor-requirements function and that its estimates are correct. Moreover, it is assumed that the firm knows exactly how today's research and investment affect its ability to produce tomorrow. These assumptions are obviously restrictive and counterfactual. They do, however, enable the focus to be placed on the stochastic aspect of regulation as opposed to the firm's imperfect knowledge of its own capabilities. Some forms of uncertainty on the firm's part could be built into the model, with the additional

they have available about the firm—its capital stock, its stock of knowledge, its operating costs, the current demand for its output, and so on—and they attempt to set a price that will make the firm's return on its *capital* stock a “fair” one in period $t + 1$.

It is important to note that the regulators are assumed to exclude the stock of knowledge from the rate base. In this model, the regulated firm's rate base consists only of its stock of physical capital. As the labor-requirements function (1) clearly indicates, technology or accumulated research plays a role in determining how much it costs the firm to produce different outputs. To this extent, the regulators take the stock of knowledge into account in setting the price the firm can charge. But in determining this price, the regulators will be trying to permit a fair return on physical capital alone, not on the combined stock of accumulated investment plus accumulated research.

In trying to anticipate what the regulators' decision will be if there is a review in period $t + 1$, the firm does have some information available. It has the record of what happened when the regulators reviewed the firm in the past. Hence, while the firm does not know precisely what rule the regulators will use, it has some evidence for constructing a mapping, from K_{t+1} and A_{t+1} to p_{t+1} , that represents the regulators' price-setting rule.⁸ Since the demand function is stable through time, the firm can use the evidence accumulated through period t to construct a regulatory output reaction function:

$$q_{t+1} = \phi'(K_{t+1}, A_{t+1}). \quad (4)$$

This function predicts what output the regulators will require the firm to produce, $q_{t+1} = p^{-1}(p_{t+1})$, if the firm is reviewed and if its capital stock and technology in period $t + 1$ are K_{t+1} and A_{t+1} , respectively. A *ceteris paribus* increase in the capital stock will affect the required output in two ways. It will increase the rate base on which the regulators are allowing the firm a fair return, and it will reduce the labor cost associated with any given output level. A *ceteris paribus* increase in the stock of knowledge affects the output the regulators will require only through the second route, namely, by reducing the labor costs for each output level. The ϕ' function is assumed to be twice continuously differentiable.

The superscript t on ϕ indicates that the function is based on the data of reviews in all periods up to and including period t .⁹ If the

probability distributions creating some notational problems, but without substantially affecting the analysis. Incorporating other elements of uncertainty—perhaps, the most important ones, as the period-by-period uncertainty about the effect of research on technology—would require much more work.

⁸ Recall from the labor-requirements function (1), that labor costs, wL , follow directly once one fixes K , A , and q (or p). In setting p (or q) to guarantee a particular rate of return on K , it follows that only knowledge of K and A and some idea of the labor-requirements function are required. Hence, in trying to determine how the regulators set p (or q), one need only observe K , A and the regulators' choice of p . The regulators' view of the firm's labor-requirements function will be one of the elements affecting the parameters of the ϕ' mapping.

⁹ One possibility, for example, is that the firm uses the observations on the regulators' past review behavior to estimate a linear regression with required output being a function of capital stock and technology. Then the firm uses the estimated parameters, call them γ_0^t , γ_1^t , γ_2^t , to predict the regulators' output reaction in case there is a review next period:

$$q_{t+1} = \gamma_0^t + \gamma_1^t K_{t+1} + \gamma_2^t A_{t+1}.$$

firm is actually reviewed in period $t + 1$, it will obtain a new observation for use in estimating its regulatory mapping for output in case of review in period $t + 2$. If, on the other hand, the firm is not reviewed at the start of period $t + 1$, its information about the regulators' review behavior will remain unchanged, and it will use the same regulatory mapping ϕ^t in choosing I_{t+1} and R_{t+1} as it is about to use in selecting I_t and R_t .¹⁰

When the firm makes its investment-research decision in period t , it wants to take account of the uncertain future ramifications of its current decision. It knows that it may be reviewed in *any* future period—not only in period $t + 1$. But, then, it must use some rule for describing the regulator's behavior in all future periods, not just in period $t + 1$. Under the assumptions of the model, the probability of a review in any future period is known to be α . As for a prediction of what the regulators will require if there is a review in some future period, the best the firm can do is use all the information it has in period t . In short, in choosing I_t and R_t the firm will assume that if it is reviewed in period $t + j$, with $j \geq 1$, then the regulators will require it to produce

$$q_{t+j} = \phi^t(K_{t+j}, A_{t+j}). \quad (5)$$

It should be emphasized that (5) represents an *assumption* the firm makes in order to effect its period t investment-research decision, and it is that decision on which we are focusing. In *real* time, however, if the firm is reviewed in $t + j$ the output the regulators will require will probably differ from $\phi^t(K_{t+j}, A_{t+j})$. Moreover, when the firm actually makes its investment and research decision in period $t + j - 1$, it will assume the regulators' reaction function to be

$$q_{t+j+k} = \phi^{t+j-1}(K_{t+j+k}, A_{t+j+k}).$$

The parameters (and form) of $\phi^{t+j-1}(\cdot)$ will be the same as that of $\phi^t(\cdot)$ if the firm is never reviewed between period t and period $t + j$. If the firm is reviewed in that time interval, the function $\phi^{t+j-1}(\cdot)$ may well differ from $\phi^t(\cdot)$.¹¹

There remains the question of how far into the future the firm will look when it makes its investment-research decision for period t . What is its horizon in deciding on today's investment and today's research? While it seems clear that the firm will not want to extrapolate its current view of the regulators' behavior for an infinite period, any choice of a finite horizon is arbitrary. It is to be hoped

¹⁰ This formulation of the firm's expectations about regulators' behavior is clearly quite simplistic. Two generalizations are worth noting. First, the firm may not have a point estimate of what the required output will be if a review occurs when the stock input levels are K_{t+1} and A_{t+1} . Instead, the required output might be a random variable depending on the state of nature. In this case, given K_{t+1} and A_{t+1} , the $t + 1$ st period output would be a random variable taking on the value q_t with probability $1 - \alpha$, and taking on other values with probabilities summing to α . Second, moving one step further back, the probability of review may not be known to the firm, and it may use its observations on when reviews occur to estimate this probability. In this case, the occurrence of a review would not only provide a new observation for use in estimating the regulatory mapping but it would also alter the firm's estimate of the probability of review itself. Furthermore, in this regime, the absence of a review in period $t + 1$ would leave the firm's estimate of the regulatory mapping unchanged, but it would affect the firm's estimate of the probability of review in period $t + 2$. The analysis of these generalizations we leave for another time.

that the finite horizon chosen is far enough into the future that the exact choice does not substantially affect the firm's period t decision. If this is the case, the arbitrary character of the choice is not so bothersome since the firm is, after all, only interested in the I_t and R_t values when it solves its period t subjective problem. The policies that emerge from that subjective problem for investment and research in periods beyond t are of no interest in themselves; they are only a means to the end of determining I_t and R_t .

It will be assumed that the firm uses a T -period horizon when it makes its investment and research decision for a given period. That is, in choosing I_t and R_t the firm supposes that the last possible review will occur in period $t + T$, and that it will do neither investing nor research after period $t + T - 1$. The firm assumes that in every period from $t + T + 1$ onward, it will have to produce the same output as it will have to produce in period $t + T$ with the capital and technology it will have in period $t + T$.

The firm's assumptions about the regulatory process when choosing its investment and research levels for period t can be conveniently summarized at this point. The firm bases its decision on the following assumption.

For $1 \leq j \leq T$,

$$q_{t+j} = \begin{cases} \phi^t(K_{t+j}, A_{t+j}) & \text{with probability } \alpha \\ q_{t+j-1} & \text{with probability } 1 - \alpha. \end{cases} \quad (6)$$

For $j \geq T + 1$,

$$q_{t+j} = q_{t+T}.$$

This may be stated somewhat more compactly if a new random variable is introduced. Let u_{t+j} be a random variable indicating whether or not there is a review in period $t + j$. It takes on the value 1 with probability α and 0 with probability $1 - \alpha$. The firm's assumption (6) can now be written as:

$$\begin{aligned} q_{t+j} &= u_{t+j}\phi^t(K_{t+j}, A_{t+j}) + (1 - u_{t+j})q_{t+j-1} & \text{for } 1 \leq j \leq T \\ q_{t+j} &= q_{t+T} & \text{for } j \geq T + 1. \end{aligned} \quad (6a)$$

One further assumption is required about the regulatory process and the nature of the firm's production-demand environment. When the regulators review the firm and force it to lower its price (equivalently, under static demand, force it to increase its output) they do so in an attempt to reduce the firm's net revenue. If such regulatory behavior is to be rational, it must be the case that *given* a capital stock

¹¹ Using the linear-regression version of $\phi^t(K_{t+j}, A_{t+j})$, in choosing I_t and R_t the firm acts under the assumption that a review in period $t + j$ will lead it to produce

$$q_{t+j} = \gamma_0^t + \gamma_1^t K_{t+j} + \gamma_2^t A_{t+j}.$$

If it is actually reviewed in $t + j$, the output it will be required to produce will probably differ from $\gamma_0^t + \gamma_1^t K_{t+j} + \gamma_2^t A_{t+j}$. And, when the firm actually makes its choice of I_{t+j-1} and R_{t+j-1} , it will assume the regulators' reaction function to be

$$q_{t+j+k} = \gamma_0^{t+j-1} + \gamma_1^{t+j-1} K_{t+j+k} + \gamma_2^{t+j-1} A_{t+j+k}.$$

The parameter γ_i^{t+j-1} will equal γ_i^t for $i = 0, 1, 2$ if no review takes place between periods t and $t + j$. But if a review takes place some time during that interval, the γ_i^{t+j-1} values will probably differ from their γ_i^t counterparts.

and a stock of knowledge, any *ceteris paribus* increase in output (decrease in price) must lower or leave unchanged the firm's net revenue. It will be assumed that this is true, so that

$$\frac{\partial \text{Net Revenue}}{\partial q} = G'(q) - \frac{\partial \mathcal{L}(q, K, A)}{\partial q} \leq 0 \quad (7)$$

for all $(q, K, A) \geq 0$.

□ **A mathematical statement of the firm's problem.** The building blocks of the model have now been presented. It is assumed that the firm's discount factor is β , with $0 < \beta < 1$, and that its goal is to maximize the expected discounted present value of its cash flow, that is, to maximize its net worth. At time t the firm solves the following *subjective* problem which yields an optimal investment-research policy in subjective time for periods t through $t + T - 1$. But subjective time t is also objective (or real) time t , and hence the firm can extract from its solution to this subjective problem the optimal levels for investment and research in period t , namely, I_t and R_t , respectively. The firm's subjective problem at time t is:

$$\begin{aligned} \text{Maximize } E\{ \sum_{j=0}^{T-1} \beta^{t+j} [G(q_{t+j}) \\ - w\mathcal{L}(q_{t+j}, K_{t+j}, A_{t+j}) - cI_{t+j} - R_{t+j}] \\ + \sum_{j=T}^{\infty} \beta^{t+j} [G(q_{t+j}) - w\mathcal{L}(q_{t+j}, K_{t+j}, A_{t+j})] \} \end{aligned}$$

Subject to: $K_{t+j+1} - K_{t+j} = I_{t+j}$ all j

$A_{t+j+1} - A_{t+j} = R_{t+j}$ all j

$$q_{t+j} = \begin{cases} \phi^t(K_{t+j}, A_{t+j}) \\ \text{with probability } \alpha \text{ (review)} \\ q_{t+j-1} \\ \text{with probability } 1 - \alpha \text{ (no review)} \end{cases} \quad (8)$$

for $1 \leq j \leq T$.

$q_{t+j} = q_{t+T}$ for $j \geq T + 1$.

$I_{t+j} \geq 0, R_{t+j} \geq 0$ for $0 \leq j \leq T - 1$;

$I_{t+j} = R_{t+j} = 0$ for $j \geq T$.

q_t, K_t, A_t given.

Several notational simplifications should be taken advantage of immediately. First, note that the form of the firm's problem is the same for all t . It suffices, then, to consider the firm's decision for time period $t = 0$. The t subscripts on variables and the t superscript on the ϕ function can therefore be omitted. Of course, when the firm moves ahead in real time, the data of the problem change accordingly—in particular, the ϕ^t function changes. The form of the problem, however, remains the same.

Second, since the expression for gross revenue minus labor costs appears so often, it pays to have a special notation for it. Hence,

define the function $N(q, K, A)$ as

$$N(q, K, A) \equiv G(q) - w\mathcal{L}(q, K, A).$$

Note that since both $G(q)$ and $\mathcal{L}(q, K, A)$ are twice continuously differentiable, so is $N(q, K, A)$. In this notation, the important assumption embodied in (7) is:

$$N_1(q, K, A) = \frac{\partial N(q, K, A)}{\partial q} \leq 0 \quad \text{for all } (q, K, A) \geq 0, \quad (7)$$

while it follows from (1) that $N_2(q, K, A) > 0$, and $N_3(q, K, A) > 0$ for all $(q, K, A) \geq 0$.

With these notational simplifications, after summing the terms from T to ∞ , the objective function in (8) becomes,

$$E \left\{ \sum_{j=0}^{T-1} \beta^j [N(q_j, K_j, A_j) - cI_j - R_j] + \frac{\beta^T}{1-\beta} N(q_T, K_T, A_T) \right\}.$$

Since q_0, K_0, A_0 are given at the outset, maximizing this function is equivalent to maximizing

$$E \left\{ \sum_{j=0}^{T-1} \beta^j [\beta N(q_{j+1}, K_{j+1}, A_{j+1}) - cI_j - R_j] + \frac{\beta^{T+1}}{1-\beta} N(q_T, K_T, A_T) \right\}.$$

Using the definition of u_t given earlier, the firm's problem at $t = 0$ can now be stated as follows:

$$\begin{aligned} \text{Maximize}_{\{I_j\}, \{R_j\}} E \left\{ \sum_{j=0}^{T-1} \beta^j [\beta N(q_{j+1}, K_{j+1}, A_{j+1}) - cI_j - R_j] \right. \\ \left. + \frac{\beta^{T+1}}{1-\beta} N(q_T, K_T, A_T) \right\} \end{aligned}$$

Subject to: $K_{j+1} - K_j = I_j \quad \text{all } j$

$$A_{j+1} - A_j = R_j \quad \text{all } j$$

$$q_j = u_j \phi(K_j, A_j) + (1 - u_j) q_{j-1} \quad \text{for } 1 \leq j \leq T \quad (9)$$

$$q_j = q_T \quad \text{for } j \geq T + 1.$$

$$I_j \geq 0, R_j \geq 0 \quad \text{for } 0 \leq j \leq T - 1;$$

$$I_j = R_j = 0 \quad \text{for } j \geq T.$$

$$q_0, K_0, A_0 \text{ given.}$$

In the next section, the optimal solution to the firm's problem, (9), is presented.

3. The firm's optimal investment and research policy

■ The techniques of dynamic programming are used to derive a characterization of the optimal policy for all j in the solution to (9). In deriving the basic functional equation, it is useful to identify $\beta N(q_{j+1}, K_{j+1}, A_{j+1}) - cI_j - R_j$ as the value of the j th "stage" of the decision process. The first term, $\beta N(q_{j+1}, K_{j+1}, A_{j+1})$, is the value, discounted to period j , of the gross revenue minus the wage bill in period $j + 1$, and this is determined by the investment-research decision in period j and the regulatory review process. From this one

subtracts the cost of the investment and research done in period j to obtain the value of the j th "stage" of the process.

Define $V_j(q, K, A)$ as the maximum expected profit, discounted to period j , from stage j onward given that the state of the system at the start of period j is $\{q, K, A\}$. That is, $V_j(q, K, A)$ is the expected value discounted to time j of an optimal policy starting at time j with the output to be produced during period j being q , the capital stock on hand at the start of period j being K , and the stock of knowledge at the start of period j being A . Stated explicitly, for $j = T$

$$V_T(q, K, A) \equiv \frac{\beta}{1 - \beta} N(q, K, A), \quad (10)$$

while for $0 \leq j \leq T - 1$, $V_j(q, K, A)$ is defined recursively by the functional equation:

$$\begin{aligned} V_j(q, K, A) = \text{Max}_{I \geq 0, R \geq 0} & E\{\beta N(u_{j+1}\phi(K + I, A + R) \\ & + (1 - u_{j+1})q_j, K + I, A + R) - cI - R \\ & + \beta V_{j+1}(u_{j+1}\phi(K + I, A + R) \\ & + (1 - u_{j+1})q_j, K + I, A + R)\}. \end{aligned} \quad (11)$$

Evaluating the expectation explicitly, this becomes:

$$\begin{aligned} V_j(q, K, A) = \text{Max}_{I \geq 0, R \geq 0} & \{\alpha\beta N[\phi(K + I, A + R), K + I, A + R] \\ & + (1 - \alpha)\beta N(q, K + I, A + R) - cI - R \\ & + \alpha\beta V_{j+1}[\phi(K + I, A + R), K + I, A + R] \\ & + (1 - \alpha)\beta V_{j+1}(q, K + I, A + R)\}. \end{aligned} \quad (11a)$$

An optimal policy at (subjective) time j , given required output q , capital stock K , and technology A at the start of period j , is comprised of the investment $I_j^*(q, K, A)$ and the research $R_j^*(q, K, A)$ which maximize the right-hand side of (11) or (11a) subject to the constraints. The objective function in that right-hand side has a straightforward interpretation. The first two terms in the maximand in (11a) represent the expected gross revenue minus labor costs for period $j + 1$ (discounted to period j) given that the investment in period j is I and the research in period j is R . The next two terms give the cost of that investment and research program. Finally, the last two terms present the discounted (to period j) expected value of an optimal policy from period $j + 1$ onward *given* the capital stock and stock of knowledge the firm inherits at the start of period $j + 1$ as a result of its period j investment-research decision.

In characterizing the optimal policy, it will be convenient to write (11a) somewhat differently. Specifically, introducing the variables $Y \equiv K + I$ and $Z \equiv A + R$, (11a) can be expressed as:

$$\begin{aligned} V_j(q, K, A) = \text{Max}_{Y \geq K, Z \geq A} & \{\alpha\beta N[\phi(Y, Z), Y, Z] + (1 - \alpha)\beta N(q, Y, Z) \\ & - c(Y - K) - (Z - A) + \alpha\beta V_{j+1}[\phi(Y, Z), Y, Z] \\ & + (1 - \alpha)\beta V_{j+1}(q, Y, Z)\}. \end{aligned} \quad (12)$$

If $Y_j^*(q, K, A)$ and $Z_j^*(q, K, A)$ are the maximizing values in (12), then $Y_j^* - K = I_j^*$ is the optimal investment in period j and $Z_j^* - A = R_j^*$ is the optimal research level for period j . Finally,

define $H_j(q, Y, Z)$ as follows:

$$H_j(q, Y, Z) \equiv -cY - Z + \alpha\beta N[\phi(Y, Z), Y, Z] \\ + (1 - \alpha)\beta N(q, Y, Z) + \alpha\beta V_{j+1}[\phi(Y, Z), Y, Z] \\ + (1 - \alpha)\beta V_{j+1}(q, Y, Z). \quad (13)$$

Then (12) can be written more compactly as:

$$V_j(q, K, A) = \text{Max}_{Y \geq K, Z \geq A} \{cK + A + H_j(q, Y, Z)\}, \quad (14)$$

and $Y_j^*(q, K, A)$, $Z_j^*(q, K, A)$ are the values of Y and Z that solve the problem on the right-hand side of (14).

It is now possible to characterize the form of the firm's optimal policy for each subjective period j and, in particular, for subjective period 0 which is also objective period 0. Much of the work required to characterize the optimal policy's structure involves proving that the function $H_j(q, Y, Z)$ is strictly concave for all $0 \leq j \leq T - 1$. First, Theorem 1 establishes the nature of the optimal policy for any period j under the assumption that $H_j(q, Y, Z)$ is strictly concave. Then Theorem 2 provides a set of conditions (sufficient but not necessary) under which $H_j(q, Y, Z)$ is strictly concave for all $0 \leq j \leq T - 1$. The proofs of both theorems are presented in the Appendix.

Four critical values Y_j^0 , Z_j^0 , Y_j' , and Z_j' are required to characterize the optimal investment-research policy $I_j^* = Y_j^* - K$, $R_j^* = Z_j^* - A$. These values are defined as the solutions to several programming problems as follows:

$$\begin{aligned} (Y_j^0, Z_j^0) \text{ solves the problem } (P) \quad & \text{Maximize } H_j(q, Y, Z) \\ & (Y, Z) \geq 0 \\ Z_j' \quad & \text{solves the problem } (P_K) \quad \text{Maximize } H_j(q, K, Z) \\ & Z \geq 0 \\ Y_j' \quad & \text{solves the problem } (P_A) \quad \text{Maximize } H_j(q, Y, A). \\ & Y \geq 0 \end{aligned} \quad (15)$$

It should be noted that the solutions to (P) , (P_K) , and (P_A) are each unique because $H_j(q, K, A)$ is, by assumption, strictly concave. Now divide Y_j , Z_j space into the following four mutually exclusive and collectively exhaustive regions:

$$\begin{aligned} \text{(I)} &= \{(Y_j, Z_j) \mid Y_j \geq K, Z_j \geq A\}; \\ \text{(II)} &= \{(Y_j, Z_j) \mid Y_j < K, Z_j \geq A\}; \\ \text{(III)} &= \{(Y_j, Z_j) \mid Y_j < K, Z_j < A\}; \\ \text{(IV)} &= \{(Y_j, Z_j) \mid Y_j \geq K, Z_j < A\}. \end{aligned} \quad (16)$$

With these definitions, the following theorem characterizes the form of the optimal policy for each $j = 0, \dots, T - 1$.

Theorem 1. If $H_j(q, Y, Z)$ is strictly concave, then the firm's optimal investment in period j is $I_j^* = Y_j^* - K$ and its optimal research in period j is $R_j^* = Z_j^* - A$, where

$$\begin{aligned} \text{(i)} \quad & (Y_j^0, Z_j^0) \in \text{I} \Rightarrow (Y_j^*, Z_j^*) = (Y_j^0, Z_j^0) \\ \text{(ii)} \quad & (Y_j^0, Z_j^0) \in \text{II} \Rightarrow (Y_j^*, Z_j^*) = (K, \text{Max}[A, Z_j']) \\ \text{(iii)} \quad & (Y_j^0, Z_j^0) \in \text{III} \Rightarrow (Y_j^*, Z_j^*) \text{ solves } \text{Max}_{(Y, Z) \in S} H_j(q, Y, Z) \\ & \text{with } S = \{(K, \text{Max}[A, Z_j']), (\text{Max}[K, Y_j'], A)\} \\ \text{(iv)} \quad & (Y_j^0, Z_j^0) \in \text{IV} \Rightarrow (Y_j^*, Z_j^*) = (\text{Max}[K, Y_j'], A). \end{aligned} \quad (17)$$

The optimal policy described in Theorem 1 has a rather straightforward interpretation. Examining the definition of $H_j(q, Y, Z)$ in (13), the first two terms represent the cost of acquiring a capital stock Y and a stock of knowledge Z *de novo* in period j . That is, they give the amount a firm just beginning with no capital and no technology would have to pay to acquire capital Y and technology Z . The next two terms represent the expected gross revenue minus labor costs for period $j + 1$ (discounted to period j) given that the firm's capital stock in $j + 1$ is Y and its stock of knowledge in $j + 1$ is Z . The last two terms give the expected value (discounted to j) of an optimal policy from $j + 1$ onward given that capital in $j + 1$ is Y and technology in $j + 1$ is Z .

The theorem states that the firm looks first at what capital-technology pair it would choose if it could start from scratch in period $j + 1$, namely, Y_j^0, Z_j^0 . If these levels of the two stocks exceed what the firm now possesses, it moves to these optimal levels by investing $I_j^* = Y_j^0 - K$ and doing research $R_j^* = Z_j^0 - A$. This is case (i). Suppose, in contrast, that while Y_j^0 is greater than the firm's current capital stock K_j , the desired stock of knowledge Z_j^0 is smaller than what the firm already possesses. Then the firm finds itself in case (iv). The theorem tells it to find a second-best capital stock, namely, the one it would want to have *given* that it must operate with technology at level A , and this capital stock is Y_j' . If Y_j' exceeds the current capital on hand, the firm adds to its plant and equipment an amount $I_j^* = Y_j' - K_j$, and it does no research, $R_j^* = 0$. If Y_j' is less than K_j , the capital the firm already owns, then the firm neither invests nor does research in period j : $I_j^* = R_j^* = 0$. A similar interpretation can be given to case (ii), and with a bit more complication, case (iii) can be interpreted analogously.

From Theorem 1 and the definition of $V_j(q, K, A)$ in (14), it follows that the valuation function $V_j(q, K, A)$ takes one of the several forms displayed in Table 1.

The characterization of the firm's optimal policy for period j pre-

TABLE 1
THE FORM OF THE VALUATION FUNCTION $V_j(q, K, A)$

CASE		$V_j(q, K, A)$
REGION I		$cK + A + H_j(q, Y_j^0, Z_j^0)$
REGION II	$Z_j' \geq A$	$cK + A + H_j(q, K, Z_j')$
	$Z_j' < A$	$cK + A + H_j(q, K, A)$
REGION III	$H_j(q, K, \text{MAX}[A, Z_j']) \geq H_j(q, \text{MAX}[K, Y_j'], A)$	
	$Z_j' \geq A$	$cK + A + H_j(q, K, Z_j')$
	$Z_j' < A$	$cK + A + H_j(q, K, A)$
	$H_j(q, \text{MAX}[K, Y_j'], A) \geq H_j(q, K, \text{MAX}[A, Z_j'])$	
	$Y_j' \geq K$	$cK + A + H_j(q, Y_j', A)$
	$Y_j' < K$	$cK + A + H_j(q, K, A)$
REGION IV	$Y_j' \geq K$	$cK + A + H_j(q, Y_j', A)$
	$Y_j' < K$	$cK + A + H_j(q, K, A)$

sented in Theorem 1 is based on the assumption that $H_j(q, K, A)$ is strictly concave. The following theorem provides conditions on the functions basic to the problem, namely, the demand function, the production function, and the firm's perception of the regulators' output reaction function, such that $H_j(q, K, A)$ is strictly concave for all $0 \leq j \leq T - 1$.

Theorem 2. If (i) $N(q, K, A)$ is strictly concave, (ii) $N_1(q, K, A) \leq 0$ for all $(q, K, A) \geq 0$, and (iii) $\phi(K, A)$ is convex, then $H_j(q, K, A)$ defined in (13) is strictly concave for all $0 \leq j \leq T - 1$.

The first condition of the theorem states that the gross revenue in any period minus the labor cost in that period is a strictly concave function of q , K , and A . Condition (ii) is simply the assumption stated in (7) concerning the nature of the firm's demand-production environment. Finally, condition (iii) requires that the output reaction function used by the firm be convex in capital and technology.

One other issue concerning the optimal policy in Theorem 1 remains to be considered. Specifically, the entire discussion has implicitly assumed that the critical values Y_j^0 , Z_j^0 , Y_j' , and Z_j' defined in (15) are positive and finite. In addition to knowing that this is indeed the case, it would be useful—in stating explicit defining conditions for these critical values—if something were known about the differentiability of the $H_j(q, Y, Z)$ and $V_j(q, K, A)$ functions. The following theorem states conditions under which the critical values used in describing the optimal policy are positive and finite, and it also provides information about the differentiability of the functions in question. The proof of the theorem involves a rather straightforward but tedious induction argument relying heavily on the information in Table 1, so it will be omitted. Notationally, $V_{j,i}(q, K, A)$ is the partial derivative of $V_j(q, K, A)$ with respect to its i th argument, and $V_{j,ik}(q, K, A)$ is the second partial derivative of $V_j(q, K, A)$ with respect to its i th and k th arguments, and similarly for $H_{j,i}(q, K, A)$ and $H_{j,ik}(q, K, A)$.

Theorem 3. If $N(q, K, A)$ and $\phi(K, A)$ are twice continuously differentiable functions satisfying the conditions in Theorem 2 and possessing the additional properties

(i) $N_1(q, K, A)$ is bounded from below;

(ii) $\phi_1(K, A)$ and $\phi_2(K, A)$ are both positive and bounded from above;

(iii) $\lim_{K \rightarrow 0^+} N_2(q, K, A) = \infty$, $\lim_{A \rightarrow 0^+} N_3(q, K, A) = \infty$,
 $\lim_{K \rightarrow 0^+} N_2[\phi(K, A), K, A] = \infty$, $\lim_{A \rightarrow 0^+} N_3[\phi(K, A), K, A] = \infty$;

and

(iv) $\lim_{K \rightarrow \infty} N_2(q, K, A) = 0$, $\lim_{A \rightarrow \infty} N_3(q, K, A) = 0$,
 $\lim_{K \rightarrow \infty} N_2[\phi(K, A), K, A] = 0$, $\lim_{A \rightarrow \infty} N_3[\phi(K, A), K, A] = 0$;

then

(a) the critical values Y_j^0 , Z_j^0 , Y_j' , Z_j' are unique, positive, and finite for all $0 \leq j \leq T - 1$;

(b) $H_j(q, K, A)$ and $V_j(q, K, A)$ are continuously differentiable

with $-\infty < V_{j,1}(q, K, A) \leq 0$, $-\infty < V_{j,2}(q, K, A) \leq c$,
and $-\infty < V_{j,3}(q, K, A) \leq 1$, for all $0 \leq j \leq T-1$;

- (c) $V_{j,11}(q, K, A) \leq 0$ and $H_{j,11}(q, K, A) \leq 0$ exist everywhere
for $0 \leq j \leq T-1$;

and

- (d) for each j , $0 \leq j \leq T-1$, $V_{j,22}$, $V_{j,33}$, $V_{j,12}$, $V_{j,13}$, $V_{j,23}$ and
 $H_{j,22}$, $H_{j,33}$, $H_{j,12}$, $H_{j,13}$, $H_{j,23}$ exist everywhere except possibly
at the points (q, Y_j^0, Z_j^0) , (q, Y_j', A) , and (q, K, Z_j') .

The conditions of Theorem 3 are easily interpretable. Condition (i) states that while the net marginal revenue of an additional unit of output, holding capital and technology fixed, is negative [condition (ii) of Theorem 2], there is a limit on how negative it can be. Similarly, condition (ii) states that the firm perceives that an increase in capital or in knowledge will lead to an increase in the output the regulators will require, but that the marginal effect such changes have on the required output is limited. Condition (iii) restricts the behavior of the net marginal revenue products of the stock of capital and the stock of knowledge as their respective levels approach zero, both when output is independent of the stocks and when output depends on these stocks through the $\phi(K, A)$ function. Finally, condition (iv) states that these same net marginal revenue products approach zero as the respective stocks get very large.

■ **Difficulties in sensitivity analysis: the nonmyopic character of the optimal policy.** The theorems of the preceding section characterize the optimal investment-research policy the firm should adopt for *subjective* periods $t+j$, for $j = 0, 1, \dots, T-1$. (The subscript t that was dropped in going from problem (8) to problem (9) has now been restored.) These results provide the optimal solution, in the form of an optimal investment-research policy, for the firm's subjective problem as seen at time t and as given in (8). From this solution policy, the firm extracts the answer of how much investment and research it should do in *real* period t (which is also subjective period t), namely, $I_t^* = I_t^*(q_t, K_t, A_t)$ and $R_t^* = R_t^*(q_t, K_t, A_t)$, respectively. The fact that the firm chooses its investment and research levels for period t in this way has implications for the efficiency of the firm's resource allocation and for the responsiveness of its behavior to changes in the regulatory environment, in particular, to changes in regulatory lag. This section explores some of these implications.

One major difficulty arises in analyzing the firm's investment-research decision for period t and the sensitivity of that decision to various parameter changes. The firm selects $I_t^* = Y_t^* - K_t$ and $R_t^* = Z_t^* - A_t$ to

$$\begin{aligned} & \text{Maximize } \{ -cY_t - Z_t + \alpha\beta N[\phi(Y_t, Z_t), Y_t, Z_t] \\ & Y_t \geq K_t, Z_t \geq A_t \\ & + (1 - \alpha)\beta N(q_t, Y_t, Z_t) + \alpha\beta V_{t+1}[\phi(Y_t, Z_t), Y_t, Z_t] \\ & + (1 - \alpha)\beta V_{t+1}(q_t, Y_t, Z_t) \}. \end{aligned} \quad (18)$$

(For notational simplicity, we continue to omit the superscript t on the ϕ function.) The firm's period t decision is clearly not myopic. To determine exactly how much investment and research to do in period

4. Some implications of the optimal policy for regulated firm behavior

t , it must know what it will do in *subjective* period $t + 1$ since $V_{t+1}(\cdot)$ enters the maximization problem. This, in turn, requires that the firm must know what it will do in subjective period $t + 2$, since $V_{t+2}(\cdot)$ enters the maximand for subjective period $t + 1$, and so on.

While dynamic programming algorithms can numerically solve the firm's problem, the problem's essential nonmyopic character makes it difficult to obtain analytical results about the structure of the period t solution that go beyond those contained in the previous section's theorems. If the firm's optimal policy could be stated in a form independent of the time period, then analysis of that policy would be easier. The inherent nonmyopic character of this problem with capital and technology accumulating, remaining productive, and affecting the required output through time precludes the derivation of such myopic rules.

To gain some insights into the nature of the firm's optimal investment-research choice for period t , the remainder of the paper will focus on one part of the K, A state space. In particular, it will be assumed that the firm's optimal policy for its period t subjective problem leads it to invest and to do research in period t and in subjective period $t + 1$ whether or not it is reviewed at the start of subjective period $t + 1$. That is, it is assumed that $Y_t^0 > K_t$, $Z_t^0 > A_t$, and $Y_{t+1}^0 > K_{t+1} = Y_t^0$, $Z_{t+1}^0 > A_{t+1} = Z_t^0$, although the values of Y_{t+1}^0 and Z_{t+1}^0 will depend on whether $q_{t+1} = q_t$ or $q_{t+1} = \phi(Y_t^0, Z_t^0)$.¹²

With the assumption that the firm invests and does research in subjective period $t + 1$, it follows from Table 1 that

$$V_{t+1}(q_{t+1}, K_{t+1}, A_{t+1}) = cK_{t+1} + A_{t+1} + H_{t+1}(q_{t+1}, Y_{t+1}^0, Z_{t+1}^0).$$

Using this in (18) and collecting terms, the firm's problem in period t can be written as:

$$\begin{aligned} \text{Maximize } \pi(q_t, Y_t, Z_t) = & \{(\beta - 1)(cY_t + Z_t) \\ & + \alpha\beta N[\phi(Y_t, Z_t), Y_t, Z_t] + (1 - \alpha)\beta N(q_t, Y_t, Z_t) \\ & + \alpha\beta H_{t+1}[\phi(Y_t, Z_t), Y_{t+1}^0, Z_{t+1}^0] \\ & + (1 - \alpha)\beta H_{t+1}(q_t, \bar{Y}_{t+1}^0, \bar{Z}_{t+1}^0)\} . \end{aligned} \quad (19)$$

The bars on Y_{t+1}^0 and Z_{t+1}^0 in the last expression are used to indicate that the optimal values of Y_{t+1} and Z_{t+1} will depend on whether $q_{t+1} = q_t$ or $q_{t+1} = \phi(Y_t, Z_t)$. Since π is simply $H(q_t, Y_t, Z_t)$ with the $V_{t+1}(\cdot)$ function taking a special form, it follows from Theorem 2 that π is strictly concave, and hence the problem in (19) is a strictly concave programming problem.

To determine the optimal values Y_t^* and Z_t^* , which have been assumed to be interior, partially differentiate $\pi(q_t, Y_t, Z_t)$ first with

¹² The discussion in Section 2 indicated that at one point in the analysis it would be helpful to think in terms of demand growing through time rather than in terms of a static demand function. The present discussion is the point to which reference was made. As indicated earlier, introducing demand growth would primarily lead to notational complications, without significantly sharpening the characterizations provided by the theorems in Section 3. For example, the net revenue function N would become time-dependent, and time would have to be incorporated as an argument in the firm's perception of the regulators' output-reaction function. Demand growth would, however, make quite plausible the assumption just invoked, namely, that the firm finds itself in Region I in both subjective period t and subjective period $t + 1$.

respect to Y_t and then with respect to Z_t , and equate each of these partial derivatives to zero. Given the strict concavity of π and the assumed existence of an interior solution, the resulting first-order conditions are necessary and sufficient for an optimum. Hence, Y_t^* and Z_t^* are the solution values for (20) and (21). (For convenience the subscripts t on Y_t and Z_t are omitted.)

$$0 = \frac{\partial \pi}{\partial Y} = (\beta - 1)c \left. \begin{aligned} &+ \alpha\beta\{N_1[\phi(Y, Z), Y, Z]\phi_1(Y, Z) + N_2[\phi(Y, Z), Y, Z]\} \\ &+ (1 - \alpha)\beta N_2(q_t, Y, Z) \\ &+ \alpha\beta H_{t+1,1}[\phi(Y, Z), Y_{t+1}^0, Z_{t+1}^0]\phi_1(Y, Z) \end{aligned} \right\} \quad (20)$$

$$0 = \frac{\partial \pi}{\partial Z} = (\beta - 1) + \alpha\beta\{N_1[\phi(Y, Z), Y, Z]\phi_2(Y, Z) \left. \begin{aligned} &+ N_3[\phi(Y, Z), Y, Z]\} + (1 - \alpha)\beta N_3(q_t, Y, Z) \\ &+ \alpha\beta H_{t+1,1}[\phi(Y, Z), Y_{t+1}^0, Z_{t+1}^0]\phi_2(Y, Z) \end{aligned} \right\} \quad (21)$$

These conditions are easily interpreted. If r is defined as the firm's discount rate so that $\beta = 1/1 + r$, then equations (20) and (21) require that

$$c = \frac{1}{r} \frac{\partial}{\partial Y} [E\{N(q_{t+1}, Y, Z) + H_{t+1}(q_{t+1}, \tilde{Y}_{t+1}^0, \tilde{Z}_{t+1}^0)\}]$$

and

$$1 = \frac{1}{r} \frac{\partial}{\partial Z} [E\{N(q_{t+1}, Y, Z) + H_{t+1}(q_{t+1}, \tilde{Y}_{t+1}^0, \tilde{Z}_{t+1}^0)\}].$$

(The tildes on Y_{t+1}^0 and Z_{t+1}^0 in these expressions are used to indicate that the optimal values of Y_{t+1} and Z_{t+1} will depend on the value of q_{t+1} . If $q_{t+1} = q_t$, then Y_{t+1}^0 and Z_{t+1}^0 will be optimal; if $q_{t+1} = \phi(Y_t, Z_t)$, then $\tilde{Y}_{t+1}^0$ and $\tilde{Z}_{t+1}^0$ will be optimal.) Recalling the definition of $N(q, Y, Z)$ and the interpretation of $H_{t+1}(q, Y, Z)$ given earlier, these conditions state that investment (research) in period t should be carried to the point where the cost of the marginal unit of investment (research) exactly equals the discounted (to period t) expected value of its stream of benefits from $t + 1$ onward. Of course, the same interpretation of equating cost to discounted expected future benefits would also apply to the equations defining an interior solution in the more general case of problem (18). The possibility of boundary solutions for Y_t^* and Z_t^* in the case of (18) would, however, entail the conversion of one or both of these equality conditions to inequalities.

□ **Resource misallocation and the regulated firm's optimal policy.** Equations (20) and (21) can be used to investigate the input efficiency of the regulated firm. In particular, they can help us examine whether the principal Averch-Johnson inefficiency result—that the regulated firm operates with a capital-labor ratio greater than the one that minimizes cost for the output produced—obtains in the present model. Care must be exercised, however, in posing these input

efficiency questions in the context of this model. The firm in this model will have to produce one of two output levels in period $t + 1$: q_t or the review-dictated output which it thinks will be $\phi(Y_t, Z_t)$. Unless (a) the firm has just been reviewed, q_t does in fact equal $\phi(K_t, A_t)$, and the firm does no investing and no research in period t so that the two possible $t + 1$ outputs are the same, namely, $q_t = \phi(Y_t, Z_t)$, or (b) the Y_t and Z_t chosen are such that it is optimal to meet the possible output change from q_t to $\phi(Y_t, Z_t)$ by varying labor alone, *ex post* the firm will be inefficient in at least one of the two regimes (review-no review) in period $t + 1$. That is, unless condition (a) or condition (b) obtains, if the firm would be efficient in producing q_t in period $t + 1$, it would necessarily be inefficient in producing $\phi(Y_t, Z_t)$ in that period, and vice versa.

Nevertheless, one can compare the input combination the firm actually chooses with the inputs it would have chosen if it were going to produce efficiently *either* the nonreview output q_t *or* what it thinks will be the review-dictated output $\phi(Y_t, Z_t)$. In the notation of the current model, the three conditions for input efficiency in producing a given output q —that is, the conditions for least-cost production of a given output q —are:

$$\begin{aligned} \text{(i)} \quad & (\beta - 1)c + \beta N_2(q, Y, Z) = 0 \quad \text{capital-labor efficiency} \\ \text{(ii)} \quad & (\beta - 1) + \beta N_3(q, Y, Z) = 0 \quad \text{knowledge-labor efficiency} \\ \text{(iii)} \quad & c = \frac{N_2(q, Y, Z)}{N_3(q, Y, Z)} \quad \text{capital-knowledge efficiency.} \end{aligned} \tag{22}$$

For example, efficient use of capital vis-à-vis labor requires $\frac{\partial L}{\partial K} = \frac{rc}{w}$

$$\text{or } -\mathcal{L}_2(q, Y, Z) = \frac{\left(\frac{1 - \beta}{\beta}\right)c}{w} \text{ or } N_2(q, Y, Z) = \left(\frac{1 - \beta}{\beta}\right)c \text{ which}$$

is equivalent to (22.i). Conditions (22.ii) and (22.iii) are derived in a similar fashion. The first two equations in (22) are the first-order conditions for the firm's problem (18) when the output in period t is q and the probability of a review is zero ($\alpha = 0$) from period $t + 1$ onward. That is, (22.i) and (22.ii) yield the optimal values for capital and technology for problem (18) with $q_t = q$ if the firm produces q from period $t + 1$ onward.

To compare the firm's actual choice of a (Y, Z) pair with the efficient choice for producing q_t , rewrite (20) and (21) as:

$$\begin{aligned} (\beta - 1)c + \beta N_2(q_t, Y, Z) = & -\alpha\beta\{N_1[\phi(Y, Z), Y, Z]\phi_1(Y, Z) \\ & + N_2[\phi(Y, Z), Y, Z] - N_2(q_t, Y, Z) \\ & + H_{t+1,1}[\phi(Y, Z), Y_{t+1}^0, Z_{t+1}^0]\phi_1(Y, Z)\} \end{aligned} \tag{23}$$

and

$$\begin{aligned} (\beta - 1) + \beta N_3(q_t, Y, Z) = & -\alpha\beta\{N_1[\phi(Y, Z), Y, Z]\phi_2(Y, Z) \\ & + N_3[\phi(Y, Z), Y, Z] - N_3(q_t, Y, Z) \\ & + H_{t+1,1}[\phi(Y, Z), Y_{t+1}^0, Z_{t+1}^0]\phi_2(Y, Z)\}. \end{aligned} \tag{24}$$

Clearly, conditions (22.i) and (22.ii) will be met with $q = q_t$ if and only if the right-hand sides of (23) and (24) are zero. On the other

hand, the A-J result, which requires $\frac{\partial L}{\partial K} < \frac{rc}{w}$, will obtain if and only if the right-hand side of (23) is negative. Finally, condition (22.iii) will be met if and only if either (a) the right-hand sides of (23) and (24) are both zero or (b) the ratio of the right-hand side of (23) to the right-hand side of (24) is equal to $N_2(q_t, Y, Z)/N_3(q_t, Y, Z)$.

What, then, is the sign of the right-hand side of (23) and what is the sign of the right-hand side of (24)? It suffices to examine (23) since the argument is analogous for (24). From the assumption in (7), $N_1[\phi(Y, Z), Y, Z] \leq 0$ while from the proof of Theorem 2, $H_{t+1,1}[\phi(Y, Z), Y_{t+1}^0, Z_{t+1}^0] \leq 0$. Taking these two inequalities together with the assumption that $\phi_1(Y, Z) > 0$, made explicit in (ii) of Theorem 3, it follows that

$$N_1[\phi(Y, Z), Y, Z]\phi_1(Y, Z) + H_{t+1,1}[\phi(Y, Z), Y_{t+1}^0, Z_{t+1}^0]\phi_1(Y, Z) \leq 0. \quad (25)$$

Since it is being assumed that $Y > K_t$ and $Z > A_t$, and since the most recent time at which q_t could have been determined was a review at the start of period t , with capital stock K_t and technology A_t , it is reasonable to assume that $\phi(Y, Z) > q_t$.

The remaining question about (23) concerns the sign of

$$N_2[\phi(Y, Z), Y, Z] - N_2(q_t, Y, Z),$$

which is, recalling the definition of $N(q, Y, Z)$, the sign of

$$\mathcal{L}_2(q_t, Y, Z) - \mathcal{L}_2[\phi(Y, Z), Y, Z].$$

The absolute value of the partial derivative $\mathcal{L}_2(q, Y, Z)$ indicates the decrease in the labor required to produce q that results from a marginal increase in the capital stock given the initial capital stock Y and technology Z . Hence, the question is whether a *ceteris paribus* marginal increase in capital reduces the labor required by a larger amount, holding the initial capital and technology fixed, when the initial output is higher or lower. The answer clearly rests with the nature of the capital-labor substitution possibilities in the firm's production function. If, for example, the firm's production function is a modified Cobb-Douglas function $q = L^{\delta_1}K^{\delta_2}A^{\delta_3}$, then

$$L = q^{1/\delta_1}K^{-\delta_2/\delta_1}A^{-\delta_3/\delta_1},$$

and $\mathcal{L}_{12}(q, K, A) < 0$ for all q, K, A so that

$$\mathcal{L}_2(q'', K, A) - \mathcal{L}_2(q', K, A) < 0$$

whenever $q'' > q'$. In this case we would have

$$N_2[\phi(Y, Z), Y, Z] - N_2(q_t, Y, Z) > 0.$$

While this case in which the net marginal revenue product of capital increases with *ceteris paribus* increases in the output produced seems most plausible, some production functions could give rise to different substitution possibilities.

Denote the difference $N_2[\phi(Y, Z), Y, Z] - N_2(q_t, Y, Z)$ by ΔN_2 . Then in conjunction with (25), it follows that if $\Delta N_2 < 0$, the right-hand side of (23) is positive. If, however, $\Delta N_2 > 0$, then the sign of the right-hand side of (23) is ambiguous, as it depends on the magnitudes of the derivatives involved. This discussion and the cor-

responding analysis of (24) can be summarized as follows, where $\Delta N_3 \equiv N_3[\phi(Y, Z), Y, Z] - N_3(q_t, Y, Z)$:

- If $\Delta N_2 \leq 0$, then $(\beta - 1)c + \beta N_2(q_t, Y, Z) \geq 0$.
- If $\Delta N_2 > 0$, then $(\beta - 1)c + \beta N_2(q_t, Y, Z)$ is ambiguous.
- If $\Delta N_3 \leq 0$, then $(\beta - 1) + \beta N_3(q_t, Y, Z) \geq 0$.
- If $\Delta N_3 > 0$, then $(\beta - 1) + \beta N_3(q_t, Y, Z)$ is ambiguous.

Hence, the regulated firm in the present model will generally not combine capital and labor *or* technology and labor in a way that minimizes the cost of producing q_t in period $t + 1$. But the direction of the inefficiency is not as clear-cut as in the A-J model. If $\Delta N_2 \leq 0$ and $\Delta N_3 \leq 0$, then the firm will, in fact, be using capital-labor and knowledge-labor ratios that are *lower* than the cost-minimizing ones. This is the direct opposite of the A-J result. In the more plausible case where $\Delta N_2 > 0$ and $\Delta N_3 > 0$, the direction of the input bias in producing q_t is not necessarily toward too much capital relative to labor, as in the A-J model, but rather the directions of the inefficiencies depend on the specific production, demand, and output-reaction functions involved. Finally, the efficiency of the capital-technology mix depends on the magnitudes of the expressions on the right-hand sides of (23) and (24).

When the firm plans its t th period investment of $Y - K_t$ and its t th period research of $Z - A_t$, using (20) and (21), it assumes that if it is reviewed at the start of period $t + 1$, it will have to produce $\phi(Y, Z)$ in period $t + 1$. Are its input choices of Y and Z the efficient ones for producing $\phi(Y, Z)$? To investigate this, compare (20) and (21) with (22.i)–(22.iii), substituting $\phi(Y, Z)$ for q in (22.i)–(22.iii). Omitting the algebraic details, the results are:

- If $\Delta N_2 < 0$, then $(\beta - 1)c + \beta N_2[\phi(Y, Z), Y, Z]$ is ambiguous.
- If $\Delta N_2 \geq 0$, then $(\beta - 1)c + \beta N_2[\phi(Y, Z), Y, Z] \geq 0$.
- If $\Delta N_3 < 0$, then $(\beta - 1) + \beta N_3[\phi(Y, Z), Y, Z]$ is ambiguous.
- If $\Delta N_3 \geq 0$, then $(\beta - 1) + \beta N_3[\phi(Y, Z), Y, Z] \geq 0$.

Thus, if in period $t + 1$ the regulated firm is told to produce what it perceived the review-dictated output would be given its period t capital-technology decision, namely, $\phi(Y, Z)$, it will generally not produce this output with a cost-minimizing mix of inputs. In the more plausible case, $\Delta N_2 \geq 0$ and $\Delta N_3 \geq 0$, where the net marginal revenue productivities of capital and knowledge each increase with *ceteris paribus* increases in the output produced, the direction of the input inefficiency is precisely the opposite of that in the A-J model. In this case, the firm under stochastic review would produce the review-dictated output with capital-labor and knowledge-labor ratios *lower* than the cost-minimizing ones.

In summary, then, no matter which of the two possible period $t + 1$ outputs is considered, q_t or $\phi(Y_t, Z_t)$, the firm's input combination with capital stock Y_t and technology Z_t will in general not be the efficient (cost-minimizing) factor choice. But the A-J result of a capital-labor ratio in excess of the cost-minimizing one for the output produced also does not necessarily obtain. Under the more plausible assumption about production capabilities, namely, $\Delta N_2 \geq 0$ and $\Delta N_3 \geq 0$, the firm would, in fact, operate with a capital-labor

ratio *lower* than the one that minimizes the cost of producing $\phi(Y_t, Z_t)$ while the way the actual ratio compares with the cost-minimizing one for output q_t depends on more specific properties of the production, demand and output-reaction functions.¹³

□ **Regulatory lag and the firm's optimal policy.** This model of the regulated firm under stochastic review also raises some questions about the effectiveness of regulatory lag in generating research and development. In this model expected regulatory lag can be identified as the expected time between reviews, $1/\alpha$. Hence, investigating how changes in lag length affect the firm's behavior is relatively straightforward. One performs a comparative-statics experiment on (20)–(21) with respect to a change in the probability of a review in period $t + 1$. Totally differentiating (20)–(21) with respect to α and using Cramer's Rule together with the strict concavity of $\pi(q, Y, Z)$, it follows that

$$\text{sign} \frac{\partial Y}{\partial \alpha} = \text{sign} \{ -\pi_{Y\alpha}\pi_{ZZ} + \pi_{Z\alpha}\pi_{YZ} \} \quad (28)$$

and

$$\text{sign} \frac{\partial Z}{\partial \alpha} = \text{sign} \{ -\pi_{Z\alpha}\pi_{YY} + \pi_{Y\alpha}\pi_{YZ} \}. \quad (29)$$

(Notationally, $\pi_{Y\alpha} = \frac{\partial^2 \pi}{\partial Y \partial \alpha}$, $\pi_{YZ} = \frac{\partial^2 \pi}{\partial Y \partial Z}$, $\pi_{YY} = \frac{\partial^2 \pi}{\partial Y^2}$, and so on.)

It should be emphasized that several important assumptions are embodied in this comparative-statics experiment on the probability of a review in $t + 1$. First, it is being assumed that the probability of all future reviews remains unchanged. Also implicit is the assumption that even though the probability of a review in $t + 1$ has changed, Y_{t+1}^0, Z_{t+1}^0 remain the optimal capital-technology levels for period $t + 2$ if there is a review at the start of period $t + 1$. That is, it has been assumed that the solution values for the maximization of $H_{t+1}[\phi(Y_t, Z_t), Y_{t+1}, Z_{t+1}]$ remain unchanged.

The argument that increasing the regulatory lag increases the amount of cost-saving technical change the firm will attempt to achieve is equivalent, in this model, to the assertion that $\frac{\partial Z}{\partial \alpha} < 0$.

For as the probability of review decreases the mean length of time between reviews increases, and hence, according to this argument, the amount of research done, R_t , will increase. Starting from the same stock of knowledge A_t , this is equivalent to saying that $\frac{\partial Z}{\partial \alpha} < 0$.

Considering the terms in (29) [the analysis for capital being analogous using (28)] the concavity of π ensures that $\pi_{YY} < 0$. In contrast, the sign of π_{YZ} does not follow directly from the concavity and convexity assumptions made in deriving the optimal policy. Neither the conditions of Theorems 2 and 3 nor the results of those theorems yield definite information about the sign of π_{YZ} . Moreover, being able to attribute a particular sign to π_{YZ} requires assumptions about more

¹³ It is worth noting that in a model which is identical with the present one except that there exists only one stock input subject to the firm's control—physical capital, K ; the stock of knowledge being fixed—these results concerning the firm's inefficient but not necessarily overly capital-intensive ($A - J$) use of inputs still obtain.

than just the single-period function $N(q, Y, Z)$, which describes the basic production-demand environment faced by the firm. To reach definite conclusions about π_{YZ} one must also make more explicit assumptions about the perceived output-reaction function $\phi(Y, Z)$, and, in particular, about the nature of the functions $N[\phi(Y, Z), Y, Z]$ and $H_{t+1}[\phi(Y, Z), Y_{t+1}^0, Z_{t+1}^0]$. For the cross-partial derivative π_{YZ} indicates how a marginal change in Y_t (Z_t) affects the expected marginal profitability of Z_t (Y_t). This depends not only on how marginal changes in capital (knowledge) in period $t + 1$ affect the expected profitability of knowledge (capital) in period $t + 1$ when output is fixed but also on how such changes affect this expected profitability when the required output is an explicit function, $\phi(Y_t, Z_t)$, of the stocks of knowledge and capital.¹⁴

Nevertheless, the fact that the sign of π_{YZ} is ambiguous in the absence of more precise assumptions is neither terribly surprising nor very bothersome. Even in much simpler production situations, the derivation of comparative-statics results requires assumptions about complement-substitute relationships among factors that do not follow from first principles of convexity and the like. What is at issue here is essentially whether for the firm making its period t investment-research decision, by maximizing (19), capital and knowledge are complements or substitutes.

If the ambiguity about π_{YZ} were the only uncertainty involved in evaluating (28) and (29), one might rest content with the ability to make statements about $\frac{\partial Y}{\partial \alpha}$ and $\frac{\partial Z}{\partial \alpha}$ conditional on the sign of π_{YZ} . The situation is, however, more complicated than that. Evaluating $\pi_{Y\alpha}$, one finds it is precisely $-\frac{1}{\alpha}$ times the right-hand side of (23), while an evaluation of $\pi_{Z\alpha}$ shows that it is $-\frac{1}{\alpha}$ times the right-hand side of (24). That is,

$$\pi_{Y\alpha} = \beta \{N_1[\phi(Y, Z), Y, Z]\phi_1(Y, Z) + \Delta N_2 + H_{t+1,1}[\phi(Y, Z), Y, Z]\phi_1(Y, Z)\} \quad (30)$$

and

$$\pi_{Z\alpha} = \beta \{N_1[\phi(Y, Z), Y, Z]\phi_2(Y, Z) + \Delta N_3 + H_{t+1,1}[\phi(Y, Z), Y, Z]\phi_2(Y, Z)\}, \quad (31)$$

where, as defined earlier, $\Delta N_2 = N_2[\phi(Y, Z), Y, Z] - N_2(q_t, Y, Z)$ and $\Delta N_3 = N_3[\phi(Y, Z), Y, Z] - N_3(q_t, Y, Z)$.

Using the inequality in (25) and equations (30) and (31), it follows that

$$\begin{aligned} \Delta N_2 \leq 0 &\Rightarrow \pi_{Y\alpha} \leq 0 \quad \text{while} \quad \Delta N_2 > 0 \Rightarrow \pi_{Y\alpha} \text{ is ambiguous.} \\ \Delta N_3 \leq 0 &\Rightarrow \pi_{Z\alpha} \leq 0 \quad \text{while} \quad \Delta N_3 > 0 \Rightarrow \pi_{Z\alpha} \text{ is ambiguous.} \end{aligned} \quad (32)$$

¹⁴ As an indication of how specific the assumptions must be to sign π_{YZ} , consider the fact that assuming the firm's production function to be the modified Cobb-Douglas form described earlier, $q = L^{\delta_1}K^{\delta_2}A^{\delta_3}$, and $\phi(Y, Z)$ to be a positive linear function of Y and Z , leaves the sign of π_{YZ} ambiguous. It is only determinate if further information is provided about the values of the parameters involved.

Hence, if the single-period marginal revenue product of knowledge decreases as output increases, so that $\Delta N_3 < 0$, then an increase in the probability of review decreases the marginal profitability of research. On the other hand, if the single-period marginal revenue product of knowledge rises with a *ceteris paribus* increase in output, which has been suggested as the more plausible case, whether the marginal profitability of research rises or falls when the expected lag length decreases depends on the magnitudes of the derivatives involved. Similar comments apply to capital and investment using the first statement in (32).

An increase in the probability of a regulatory review—a decrease in the expected length of time between reviews—necessarily decreases the expected profit of the firm in this model. An increase in α increases the likelihood that the firm will have to produce a larger output, $\phi(Y, Z)$ rather than q , and from assumption (7) and Theorem 2, it is known that both $N_1(q, Y, Z) \leq 0$ and $V_{t+1,1}(q, Y, Z) \leq 0$. Hence, the decrease in the regulatory lag makes the firm a less profitable venture (in an expected value sense). An increase in α has an ambiguous effect on the marginal profitability of research when ΔN_3 is positive because with $\Delta N_3 > 0$ the increase in α has two opposing influences.¹⁶ First, since the review-dictated output rises when the stock of knowledge rises, and since an output increase results in decreased expected profit, an increase in the likelihood that a review will occur raises the penalty associated with an increase in knowledge. It, therefore, tends to lower the marginal profitability of research. At the same time, however, if the single-period marginal revenue product of knowledge rises with output, so that $\Delta N_3 > 0$, the increase in the probability that a larger output will be required, which results from the increase in α , raises the marginal profitability of research. The question is which of these two opposing forces dominates, and this question cannot be answered without more specific information about the values of the derivatives in (31).

It should be clear that without making more specific assumptions than those already made about the demand, production, and regulatory output-reaction functions, it is not possible to obtain the signs

of $\frac{\partial Y}{\partial \alpha}$ and $\frac{\partial Z}{\partial \alpha}$. The investment-research decision of the regulated firm

in the present model is affected by the probability that a regulatory review will occur: the expressions in (28) and (29) will not, in general, be zero. But the way in which the firm's investment and research undertakings respond to changes in this probability (that is, to changes in the length of the regulatory lag) depends on the nature of the production, demand, and regulatory output-reaction functions. Moreover, the nature of the response depends on characteristics of these functions that go beyond the "general" assumptions economists are wont to make in analyzing production decisions. Hence, while there clearly is scope for using regulatory lag as a conscious instrument of regulatory policy to influence the regulated firm's efforts in improving its technology, it is an instrument that can only be used with careful attention being paid to the specific nature of the firm's environment.

¹⁶ The same discussion applies to the case of capital with $\Delta N_2 > 0$.

5. Concluding comments

■ This paper has presented a different view of the decision process of the regulated firm. Although numerous—and sometimes large—gaps still exist between the model presented here and the actual regulatory process, some of the shortfalls of the more standard Averch-Johnson model have been alleviated. The optimal pattern of behavior was derived for the regulated firm as it perceives itself operating in the regulatory framework presented here. Finally, the paper examined some of the implications of this optimal policy for two traditional issues of regulatory economics: the input efficiency of regulated firms and the effect of regulatory lag on research and development. In both instances, the results for the model presented here raised some questions about the conclusions of the more traditional inquiries into these matters. The examination of these two issues was, however, limited because the firm's optimal investment-research policy in this setting is nonmyopic. Another path one might profitably pursue to get further insights into the nature of the firm's optimal policy is a numerical sensitivity analysis of that policy. By assuming specific forms for the firm's production, demand, and regulatory output-reaction functions, explicit solutions could be derived for the firm's optimal behavior, and the sensitivity of these policies to selected parameter changes could be examined.

There are, however, two other tasks that probably have priority over more detailed sensitivity analyses of the firm's optimal policy. The first of these entails the development of a method for observing the system evolve over *real* time. The results in the present paper concern the firm's optimal strategy at *real* time t and *subjective* times $t + 1, t + 2, \dots, t + T - 1$. What one now wants to do is watch the system unfold as real time $t + 1$ arrives, the regulators either leave the firm's price unchanged or reset the price, and the firm proceeds to use the model in this paper to decide on its real time $t + 1$ investment and research. This requires a model of regulatory commission behavior and some type of simulation procedure for studying the complete system encompassing both the regulators and the firm they regulate.

The other important task stems from the recognition that the probability of review, or the length of regulatory lag, does affect firm behavior. Conscious use of the probability of review as an instrument of regulatory control requires, as Section 4 observed, careful attention to the specific characteristics of the firm being regulated. Nevertheless, this is an instrument of which regulatory commissions may want to avail themselves. But, then, the question is how should the probability of review, or the regulatory lag, be chosen? The answer would seem to be that the regulatory lag length should be considered another instrument of control, alongside the fair rate of return, and the two instruments should be chosen simultaneously in a way that maximizes social welfare. Essentially, what is envisioned is embedding the firm behavior derived in the present paper as a constraint in a social-welfare-maximization problem used to choose the fair rate of return and the probability of review. This would be a generalization of the problem Klevorick and Sheshinski¹⁶ solved in applying the results of the A-J model to the choice of an optimal value for the fair rate of return. It is a task that is more easily described than executed, but one that is worth pursuing.

¹⁶ In [6] and [7], respectively.

Appendix

■ **Theorem 1.** If $H_j(q, Y, Z)$ is strictly concave, then the firm's optimal investment in period j is $I_j^* = Y_j^* - K$ and its optimal research in period j is $R_j^* = Z_j^* - A$, where

- (i) $(Y_j^0, Z_j^0) \in I \Rightarrow (Y_j^*, Z_j^*) = (Y_j^0, Z_j^0)$
- (ii) $(Y_j^0, Z_j^0) \in II \Rightarrow (Y_j^*, Z_j^*) = (K, \text{Max}[A, Z_j'])$
- (iii) $(Y_j^0, Z_j^0) \in III \Rightarrow (Y_j^*, Z_j^*)$ solves $\text{Max}_{(Y, Z) \in S} H_j(q, Y, Z)$ (17)
with
 $S = \{(K, \text{Max}[A, Z_j']), (\text{Max}[K, Y_j'], A)\}$
- (iv) $(Y_j^0, Z_j^0) \in IV \Rightarrow (Y_j^*, Z_j^*) = (\text{Max}[K, Y_j'], A)$.

Proof: Since K and A are fixed in the right-hand side of (14), it follows that $V_j(q, K, A) = cK + A + H_j(q, Y_j^*, Z_j^*)$ if and only if $H_j(q, Y_j^*, Z_j^*) = \text{Max}_{Y \geq K, Z \geq A} H_j(q, Y, Z)$. Hence, it suffices to show

that the solutions (Y_j^*, Z_j^*) presented in (17) for the different cases are the solutions of problem (Q):

$$\text{Maximize } H_j(q, Y, Z) \quad (Q) \\ Y \geq K, Z \geq A$$

for the alternative locations of (Y_j^0, Z_j^0) in Y, Z space. It is also convenient, at this point, to drop the j subscripts for the remainder of the proof.

Note first that (Y, Z) feasible for (Q) $\Rightarrow (Y, Z)$ feasible for (P)

$\therefore H(q, Y^*, Z^*) \leq H(q, Y^0, Z^0)$, and hence if (Y^0, Z^0) is feasible for (Q), it solves (Q).

$\therefore (Y^0, Z^0) \in I \Rightarrow (Y^*, Z^*) = (Y^0, Z^0)$, proving (i).

Suppose (Y^0, Z^0) is not feasible for (Q). Then $(Y^0, Z^0) \in \bigcup_{II \leq k \leq IV}$ (Region k). Consider case (ii) where $(Y^0, Z^0) \in II: Y^0 < K, Z^0 \geq A$. First, we show that $Y^* = K$ in this case. Since it must be true that $Y^* \geq K$, if $Y^* \neq K$, the only other possibility is that $Y^* > K$. But suppose $Y^* > K, Z^* \geq A$ and $(Y^0, Z^0) \in II$. Then consider the point $[Y(\lambda), Z(\lambda)] = [\lambda Y^0 + (1 - \lambda)Y^*, \lambda Z^0 + (1 - \lambda)Z^*]$. Clearly for λ positive but close enough to 0, $[Y(\lambda), Z(\lambda)] \geq [K, A]$ so that for $\lambda > 0$ but close enough to 0, $[Y(\lambda), Z(\lambda)]$ is feasible for (Q). But since $H(q, Y, Z)$ is strictly concave, it is also true that for all $0 < \lambda < 1$,

$$H(q, Y(\lambda), Z(\lambda)) > \lambda H(q, Y^0, Z^0) + (1 - \lambda)H(q, Y^*, Z^*) \\ \geq \min[H(q, Y^0, Z^0), H(q, Y^*, Z^*)] = H(q, Y^*, Z^*),$$

since Y^0, Z^0 solves the less constrained problem (P). But, then, for $\lambda > 0$ but close enough to zero $[Y(\lambda), Z(\lambda)]$ is feasible for (Q) and $H(q, Y(\lambda), Z(\lambda)) > H(q, Y^*, Z^*)$. This is a contradiction since by hypothesis Y^*, Z^* solves (Q). Therefore, $(Y^0, Z^0) \in II \Rightarrow Y^* = K$. But, then we seek Z that solves the problem $\text{Maximize } H(q, K, Z)$.

Since $H(q, Y, Z)$ is strictly concave, it follows that $Z^* = Z'$ if $Z' \geq A$ and $Z^* = A$ otherwise. Thus, we have shown that

$$(Y^0, Z^0) \in II \Rightarrow (Y^*, Z^*) = (K, \text{Max}[A, Z']),$$

proving (ii).

FIGURE A1

A REGION III CASE WITH
 $Y^* > K, Z^* = A$

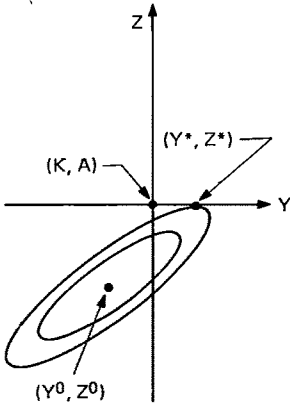
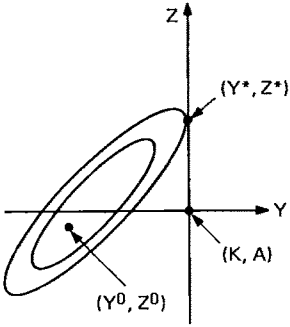


FIGURE A2

A REGION III CASE WITH
 $Y^* = K, Z^* > A$



The proof of (iv) is completely analogous to the proof of (ii). Finally, then, consider case (iii): $Y^0 < K, Z^0 < A$. Using an argument completely analogous to the one used in proving $Y^* = K$ in case (ii), it can be shown that if $(Y^0, Z^0) \in \text{III}$, then $Y^* = K$ or $Z^* = A$. That is, it cannot be true that $Y^* > K$ and $Z^* > A$. But, then, either $Y^* = K$ and Z^* maximizes $H(q, K, Z)$ subject to the constraints $Z \geq A$ or $Z^* = A$ and Y^* maximizes $H(q, Y, A)$ subject to the constraint $Y \geq K$. In the first instance, the solution values are $(K, \text{Max}[A, Z'])$, and in the second case they are $(\text{Max}[K, Y'], A)$. The firm wants to choose as (Y^*, Z^*) the member of this pair of possible solutions that yields the maximum value for $H(q, Y, Z)$. This is precisely what case (iii) of the theorem states. Figures (A1) and (A2) illustrate two possibilities for Region III. The proof of the theorem is complete.

□ *Theorem 2.* If (i) $N(q, K, A)$ is strictly concave, (ii) $N_1(q, K, A) \leq 0$ for all $(q, K, A) \geq 0$, and (iii) $\phi(K, A)$ is convex, then $H_j(q, K, A)$ defined in (13) is strictly concave for all $0 \leq j \leq T - 1$.

The proof of the theorem uses three well-known results which, for completeness, we state (without proofs) as lemmas.

Lemma 1. Given the maximization problem

$$\text{Maximize } f(x, y)$$

$$\text{Subject to: } g(x) \geq b,$$

where x is an n -vector, y is an r -vector, $f(x, y)$ is a strictly concave real-valued function, $g(x)$ is an m -vector-valued concave function and b is an m -vector, the function $\psi(b, y)$ defined by

$$\psi(b, y) \equiv \text{Max}_{x \ni g(x) \geq b} f(x, y)$$

is a strictly concave function of b and y .

Remark. For the special case in which $y = b$ (so that $r = m$), the result of Lemma 1 is that the function $\psi(b)$ is strictly concave where $\psi(b) \equiv \text{Max}_{x \ni g(x) \geq b} f(x, b)$.

Lemma 2. Let x_1, x_2, x_3, x_4, x_5 be real-valued variables and let $u(x_1, x_2, x_3)$ and $w(x_4, x_5)$ be real-valued functions. If (i) $u(x_1, x_2, x_3)$ is strictly concave, (ii) $u(x_1, x_2, x_3)$ is everywhere nonincreasing in x_1 for x_2, x_3 fixed, and (iii) $w(x_4, x_5)$ is convex, then $u(w(x_4, x_5), x_2, x_3)$ is a strictly concave function of x_2, x_3, x_4 , and x_5 .

Lemma 3. Let x_1, x_2, x_3 be real-valued variables, and let $u(x_1, x_2, x_3)$ and $w(x_2, x_3)$ be real-valued functions. If (i) $u(x_1, x_2, x_3)$ is strictly concave, (ii) $u(x_1, x_2, x_3)$ is everywhere nonincreasing in x_1 for x_2, x_3 fixed, and (iii) $w(x_2, x_3)$ is convex, then $u[w(x_2, x_3), x_2, x_3]$ is a strictly concave function of x_2 and x_3 .

With these preliminary statements of well-known results in hand, the proof of Theorem 2 can be presented.

Proof of Theorem 2. The proof is by induction on j , working backwards from $j = T - 1$ to $j = 0$. For each j , the following four pro-

positions are proven:

- (a) $H_j(q, K, A)$ is strictly concave in q, K, A .
- (b) $V_j(q, K, A)$ is strictly concave in q, K, A .
- (c) $V_j(q, K, A)$ is everywhere nonincreasing in q for K, A fixed.
- (d) $V_j(\phi(K, A), K, A)$ is strictly concave in K and A .

$j = T - 1$

From (10), $V_T(q, K, A) = \frac{\beta}{1 - \beta} N(q, K, A)$. Using this in the definition of $H_j(q, Y, Z)$ in (13) with $j = T - 1$, it follows, after combining terms, that

$$H_{T-1}(q, K, A) = -cK - A + \frac{\alpha\beta}{1 - \beta} N(\phi(K, A), K, A) + \frac{(1 - \alpha)\beta}{1 - \beta} N(q, K, A). \quad (A1)$$

From conditions (i)–(iii) of the theorem, $N(q, K, A)$ is strictly concave, and the functions $N(q, K, A)$ and $\phi(K, A)$ satisfy the conditions of Lemma 3. Thus, it follows from the lemma that $N[\phi(K, A), K, A]$ is strictly concave in K and A . But with $0 < \alpha < 1$, $\frac{\alpha\beta}{1 - \beta}$ and $\frac{(1 - \alpha)\beta}{1 - \beta}$ are both positive. Hence, $H_{T-1}(q, K, A)$ is the sum of a function linear in K and A and a positive combination of strictly concave functions of q, K , and A . Therefore,

$$H_{T-1}(q, K, A) \text{ is strictly concave in } q, K, A. \quad (A2)$$

From (14), $V_{T-1}(q, K, A) = cK + A + \max_{Y \geq K, Z \geq A} H_{T-1}(q, Y, Z)$.

But from (A2) and the linearity of the constraints, the problem

$\max_{Y \geq K, Z \geq A} H_{T-1}(q, Y, Z)$ meets the conditions of Lemma 1. Hence

$\max_{Y \geq K, Z \geq A} H_{T-1}(q, Y, Z)$ is a strictly concave function of q, K, A .

Thus, $V_{T-1}(q, K, A)$ is the sum of a function linear in K and A and a strictly concave function of q, K, A so that

$$V_{T-1}(q, K, A) \text{ is strictly concave in } q, K, A. \quad (A3)$$

Differentiating (A1) partially with respect to q , one obtains $\frac{\partial H_{T-1}(q, K, A)}{\partial q} = \frac{(1 - \alpha)\beta}{1 - \beta} \frac{\partial N(q, K, A)}{\partial q} \leq 0$ by condition (ii) of the theorem. Hence,

$$H_{T-1}(q, Y, Z) \text{ is everywhere nonincreasing in } q \text{ for } Y, Z \text{ fixed.} \quad (A4)$$

Suppose $V_{T-1}(q, K, A)$ is not everywhere nonincreasing in q when K and A are fixed. From the definition of $V_{T-1}(q, K, A)$, this is equivalent to the statement that $\max_{Y \geq K, Z \geq A} H_{T-1}(q, Y, Z)$ is not everywhere nonincreasing in q when K and A are fixed. But, then, there exist two triples $(\hat{q}, \hat{Y}, \hat{Z})$ and $(\bar{q}, \bar{Y}, \bar{Z})$ such that $\hat{q} < \bar{q}$,

$$\max_{Y \geq K, Z \geq A} H_{T-1}(\hat{q}, Y, Z) = H_{T-1}(\hat{q}, \hat{Y}, \hat{Z}),$$

$$\max_{Y \geq K, Z \geq A} H_{T-1}(\bar{q}, Y, Z) = H_{T-1}(\bar{q}, \bar{Y}, \bar{Z}),$$

and $H_{T-1}(\hat{q}, \hat{Y}, \hat{Z}) < H_{T-1}(\bar{q}, \bar{Y}, \bar{Z})$. Since $H_{T-1}(q, Y, Z)$ is everywhere nonincreasing in q for Y, Z fixed, this implies that

$$H_{T-1}(\hat{q}, \hat{Y}, \hat{Z}) < H_{T-1}(\bar{q}, \bar{Y}, \bar{Z}) \leq H_{T-1}(\hat{q}, \bar{Y}, \bar{Z}).$$

This, however, contradicts the fact that $\hat{Y}, \hat{Z}$ solves

$$\text{Max}_{Y \geq K, Z \geq A} H_{T-1}(\hat{q}, Y, Z).$$

Therefore, the initial assumption was false, and

$$V_{T-1}(q, K, A) \text{ is everywhere nonincreasing in } q \text{ for } K, A \text{ fixed.} \quad (\text{A5})$$

Substituting $T - 1$ for j and $\phi(K, A)$ for q in (14), one obtains:

$$\begin{aligned} V_{T-1}(\phi(K, A), K, A) \\ = \text{Max}_{Y \geq K, Z \geq A} \{cK + A + H_{T-1}(\phi(K, A), Y, Z)\}. \end{aligned} \quad (\text{A6})$$

From (A2), (A4), and condition (iii) of the theorem, it follows that the functions $H_{T-1}(q, Y, Z)$ and $\phi(K, A)$ meet the conditions of Lemma 2. Hence, $H_{T-1}[\phi(K, A), Y, Z]$ is a strictly concave function of K, A, Y , and Z . But, then, since the constraints are linear, the problem $\text{Max}_{Y \geq K, Z \geq A} H_{T-1}[\phi(K, A), Y, Z]$ satisfies the conditions of the Remark following Lemma 1. As a result,

$$\text{Max}_{Y \geq K, Z \geq A} H_{T-1}[\phi(K, A), Y, Z]$$

is a strictly concave function of K and A . From (A6),

$$V_{T-1}[\phi(K, A), K, A] = cK + A + \text{Max}_{Y \geq K, Z \geq A} H_{T-1}(\phi(K, A), Y, Z).$$

Thus, $V_{T-1}[\phi(K, A), K, A]$ is the sum of a function linear in K and A and a strictly concave function of K and A so that

$$V_{T-1}(\phi(K, A), K, A) \text{ is strictly concave in } K \text{ and } A. \quad (\text{A7})$$

Statements (A2), (A3), (A5), and (A7) establish (a)–(d) for $j = T - 1$.

Now assume that (a)–(d) have been proven for $j + 1, j + 2, \dots, T - 1$, and prove that they are true for j .

From (13), $H_j(q, K, A)$ is defined as follows:

$$\begin{aligned} H_j(q, K, A) \\ = -cK - A + \alpha\beta[N(\phi(K, A), K, A) + V_{j+1}(\phi(K, A), K, A)] \\ + (1 - \alpha)\beta[N(q, K, A) + V_{j+1}(q, K, A)]. \end{aligned} \quad (\text{A8})$$

As proven en route to (A2), $N[\phi(K, A), K, A]$ is strictly concave in K and A . From (d) for $j + 1$, $V_{j+1}[\phi(K, A), K, A]$ is strictly concave in K and A . From condition (i) of the theorem, $N(q, K, A)$ is strictly concave in q, K, A , while it follows from (b) for $j + 1$ that $V_{j+1}(q, K, A)$ is strictly concave in q, K, A . Since with $0 < \alpha < 1$, $\alpha\beta$ and $(1 - \alpha)\beta$ are positive, it follows that $H_j(q, K, A)$ equals a linear function of K and A plus a positive combination of strictly concave functions of q, K, A . Therefore,

$$H_j(q, K, A) \text{ is strictly concave in } q, K, A. \quad (\text{A9})$$

From (14), $V_j(q, K, A) = cK + A + \text{Max}_{Y \geq K, Z \geq A} H_j(q, Y, Z)$. But from (A9) and the linearity of the constraints, the maximization

problem on the right-hand side satisfies the conditions of Lemma 1. Hence, $\text{Max}_{Y \geq K, Z \geq A} H_j(q, Y, Z)$ is strictly concave in q, K, A , and it follows that

$$V_j(q, K, A) \text{ is strictly concave in } q, K, A. \quad (\text{A10})$$

Using the definition in (13), consider $H_j(q'', Y, Z) - H_j(q', Y, Z)$ for $q'' > q'$:

$$\begin{aligned} H_j(q'', Y, Z) - H_j(q', Y, Z) &= (1 - \alpha)\beta[N(q'', Y, Z) + V_{j+1}(q'', Y, Z) \\ &\quad - N(q', Y, Z) - V_{j+1}(q', Y, Z)] \\ &= (1 - \alpha)\beta[\{N(q'', Y, Z) - N(q', Y, Z)\} \\ &\quad + \{V_{j+1}(q'', Y, Z) - V_{j+1}(q', Y, Z)\}]. \end{aligned}$$

For $q'' > q'$, it follows from condition (ii) of the theorem that $N(q'', Y, Z) - N(q', Y, Z) \leq 0$, while from (c) for $j + 1$ it follows that $V_{j+1}(q'', Y, Z) - V_{j+1}(q', Y, Z) \leq 0$. Taking these together, and noting that $(1 - \alpha)\beta$ is nonnegative, we have that $q'' > q'$ implies $H_j(q'', Y, Z) \leq H_j(q', Y, Z)$. Therefore,

$$H_j(q, Y, Z) \text{ is everywhere nonincreasing in } q \text{ for } Y, Z \text{ fixed.} \quad (\text{A11})$$

Using the same argument used to prove (A5) from (A4), one moves from (A11) to the conclusion:

$$V_j(q, K, A) \text{ is everywhere nonincreasing in } q \text{ for } K, A \text{ fixed.} \quad (\text{A12})$$

Finally, consider $V_j[\phi(K, A), K, A]$. Substituting $\phi(K, A)$ for q in (14), one obtains

$$\begin{aligned} V_j[\phi(K, A), K, A] &= \text{Max}_{Y \geq K, Z \geq A} \{cK + A + H_j[\phi(K, A), Y, Z]\}. \quad (\text{A13}) \end{aligned}$$

Using (A9), (A11) and condition (iii) of the theorem, and Lemma 2 it follows that $H_j[\phi(K, A), Y, Z]$ is strictly concave in K, A, Y, Z . Application of the Remark following Lemma 1 then leads to the conclusion that $\text{Max}_{Y \geq K, Z \geq A} H_j[\phi(K, A), Y, Z]$ is strictly concave function of K and A . But since (A13) shows that $V_j(q, K, A)$ is the sum of a linear function of K and A plus

$$\text{Max}_{Y \geq K, Z \geq A} H_j[\phi(K, A), Y, Z],$$

it follows that

$$V_j(\phi(K, A), K, A) \text{ is a strictly concave function of } K \text{ and } A. \quad (\text{A14})$$

Statements (A9), (A10), (A12), and (A14) establish (a)–(d) for j given that they are true for $j + 1$.

The proof of Theorem 2 by induction is now complete.

References

1. AVERCH, H. AND JOHNSON, L. L. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol. 52, No. 5 (December 1962), pp. 1053–1069.
2. BAILEY, E. E. "Innovation and Regulatory Lag." Unpublished manuscript, presented at the Dartmouth College Seminar, "Problems of Regulation and Public Utilities," August 1970.

3. ——— AND COLEMAN, R. D. "The Effect of Lagged Regulation in an Averch-Johnson Model." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 278–292.
4. BAUMOL, W. J. AND KLEVORICK, A. K. "Input Choices and Rate-of-Return Regulation: An Overview of the Discussion." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 162–190.
5. KLEVORICK, A. K. "The Graduated Fair Return: A Further Comment." *The American Economic Review*, Vol. 61, No. 4 (September 1971), pp. 727–729.
6. ———. "The 'Optimal' Fair Rate of Return." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 122–153.
7. SHESHINSKI, E. "Welfare Aspects of a Regulatory Constraint: Note." *The American Economic Review*, Vol. 61, No. 1 (March 1971), pp. 175–178.
8. WICHES, C. R. "The Graduated Fair Return: Comment." *The American Economic Review*, Vol. 61, No. 4 (September 1971), pp. 725–726.

The telephone system as a public good: static and dynamic aspects

Roland Artle

Professor of Applied Economics
University of California, Berkeley

and

Christian Averous

Doctoral Candidate
University of California, Berkeley

The paper uses standard concepts of equilibrium and efficiency to analyze certain properties of the telephone system. The property of access/no access plays a central role in the analysis, and it is suggested that many so-called "systems" in modern life have similar access/no access properties.

The fundamental discreteness associated with access/no access is handled first in a static resource allocation framework. Also, the concept of transaction cost is brought into the analysis in relation to time budgets. It is shown that the conventional efficiency conditions of the theory of public goods do not hold. Alternative results are given and interpreted.

A dynamic analysis of the demand for telephone service is then developed. Using a set of very weak assumptions, we prove, by induction, the existence of a self-sustaining growth process in demand. The underlying assumptions are weak in the sense that they imply stationary conditions: a constant population, with constant incomes, a stationary income distribution, and no change in relative prices. The growth process which nevertheless arises can be explained jointly by the distribution of income and by the public-good property inherent in the telephone system.

■ "What are we left with? Two poles and a continuum in between? No. With a knife-edge pole of the private-good case, and with *all* the rest of the world in the public-good domain by virtue of involving some consumption externality."—Paul Samuelson.

This paper presents a theoretical analysis of the telephone system. It searches out the public-good dimensions of the system and analyzes

1. Introduction

Roland Artle received his degrees from the University of Oxford (Statistics) and the Stockholm School of Economics (Ph.D., Economics). From July 1, 1965, until January 1, 1969, he held an endowed chair in Economics in Sweden. He has had visiting appointments, in Economics, at Stanford University, the University of Hawaii, and at Harvard University. Currently he is doing research on efficiency and equity aspects of spatial systems.

Christian Averous graduated from the University of Paris (Economics) and from the Ecole Nationale des Ponts et Chaussees, Paris. He is presently a doctoral candidate in City and Regional Planning, at the University of California, Berkeley, where he is studying problems of urban transportation.

The authors have benefited from the Urban Economic Systems seminar, in the Department of Electrical Engineering, at the University of California, Berkeley, and especially from the discussions with Pravin Varaiya.

important implications of these. Specifically, the purposes of the paper are fourfold. First, we wish to show that, through its provision of *access*, the telephone system possesses the essential property of a public good.¹ Second, we suggest that the savings in transaction costs (incurred in information exchange, etc.) that the telephone system permits, are also public-good flavored. Third, we contribute a theoretical explanation for the growth in the demand for telephone service over time, which is so well established in all empirical studies. This explanation hinges crucially on a dynamic concept of public goods. Fourth, and finally, our analytical interest in the telephone system rests on a contention that insights into its functioning can enhance our understanding of other “systems,” which—due, say, to interdependencies and to nonconvexities—have proven largely intractable in theoretical work so far. In other words, we shall use the telephone system as a paradigm. We believe that many phenomena in modern life, and perhaps especially urban phenomena and systems, have characteristics similar to those of the telephone system, albeit much more complex. Unquestionably, the telephone system is a highly complicated system, in its own right, and for the purposes of the theoretical analyses to follow, we shall make several simplifying assumptions about this system.

Prima facie, it may seem quite straightforward to reach the first-mentioned aim of the paper, namely to demonstrate that the potential use of—the access to—the telephone system is a public good. All we would have to do—it seems—would be to set forth a set of conditions which would insure that a particular resource allocation, specifying use of the telephone system, is Pareto efficient.² If, then, access to the telephone system did indeed have the public-good property claimed for it, this would show up in the efficiency condition in the usual manner as a *summation* over marginal rates of substitution.

However, it turns out that there is a fundamental obstacle to such a procedure in the present problem. A discreteness, or (0,1) property, which is associated with access/no access, enters the problem in a crucial way. In turn, this means that the standard procedure used for deriving the efficiency conditions, being based on differential calculus, does not work. Referring to this fundamental difficulty in a recent paper, Paul Samuelson says that “the resulting search problem can be a formidable combinatorial task.”³

In the next section of our paper, we shall show in greater detail how this analytical problem arises. We shall also demonstrate that the problem can be reformulated in a way which will permit us to apply standard methods and to solve the problem. Thus, the result of the analysis in the next section will indeed show that access to the telephone system is a public good.

2. A static analysis of the telephone system

■ **Modeling resource allocation.** Consider a population, or group, consisting of h individuals. They are a “group” in the sense that they all enjoy talking to each other. We shall assume that each individual

¹ The concept of “public good” is well defined in [8] and [9].

² By a Pareto-efficient allocation of resources we mean a *vector* maximum over a set of utility functions.

³ See [9].

has conversation with every other individual, over a period of time (say, a week, a month, or a year), either by way of telephone or through face-to-face contact. For simplicity in the analysis, we assume that a telephone conversation is a perfect substitute for a face-to-face conversation.

For the purpose of this paper, we shall also assume that if an individual is *not* connected to the telephone system—by virtue of being a telephone subscriber—he has no access to the system at all. In this way, we can define two subsets of individuals, called G_1 and G_0 to be mutually exclusive, and such that they jointly exhaust the total number of individuals in the population h . One group, G_1 , consists of all individuals who have access to the system, and the other group, G_0 , consists of all individuals who do not have access. We can then model each individual's preference or utility function:

$$\begin{aligned} U^i(C^i, x^i), \quad x^i = x \quad \text{for all } i \text{ in } G_1, \\ x^i = 0 \quad \text{for all } i \text{ in } G_0. \end{aligned} \quad (1)$$

Here, U^i stands for the i th individual's utility function; x^i stands for the i th individual's access to the telephone system—denoted by x —or no access—denoted by 0; C^i denotes the i th individual's consumption of all other goods and services, for simplicity taken to be private.

This property of access or no access can also be symbolized through the use of dummy variables δ^i , with $\delta^i = 1$ for an individual who has access and with $\delta^i = 0$ for an individual who does not have access. An alternative formulation of the utility functions is therefore:

$$\begin{aligned} U^i(C^i, \delta^i x^i), \quad \text{with } x^i = x, \quad \text{for all } i \quad \text{and} \quad \delta^i = 1 \text{ for } i \text{ in } G_1 \\ \delta^i = 0 \text{ for } i \text{ in } G_0. \end{aligned} \quad (2)$$

The concept of preference function used here differs from the standard form where *all* individuals have access to the public good. No manipulation of the notation in (2) permits restatement of the preference function into the standard form, or vice versa.

Since our purpose is to expose in the simplest possible terms the properties of the access variable, we shall assume that all individuals connected by telephone—the “members of the telephone club,” if you wish—consume identical bundles of goods. Similarly, the individuals who do not have access to the telephone system, are assumed to be all “alike” in the sense of having identical bundles of goods. In this way, we can say that each individual with access to the telephone system has a utility function U^1 , and each individual without telephone has a utility function U^0 . Here, U^1 and U^0 can very well be the same function, or they can be different functions.⁴ The main point is that they differ in the argument values; the bundles of goods differ between them.

Thus, each individual has a utility function with the arguments consisting of telephone service (or the lack of it) and all other goods and services, as stated in (1) and (2), above.

Our next step is to employ Paul Samuelson's omniscient planner. We equip him with a social welfare function, in order to embrace the well-being of all the h individuals in the population. To make precise the technological constraint imposed on the group of h , we also

⁴ The following analysis treats U^1 and U^0 as the same function.

provide him with a social-production-possibility frontier. We are then ready to pose an interesting question, neatly embedded in a general equilibrium framework: What is the optimal number of telephones?

Now, with appropriately specified and well-behaved functions, we might indeed come up with such a number. Even if we do not, the exercise would still be rated a success, provided that it would shed enough light on the properties of having access to the telephone system.

Even in these simple terms, and as indicated in a previous section, the problem does contain a serious analytical difficulty, because of the discreteness property associated with access/no access. The key to solution lies in the formulation of the optimization problem, which we shall present in (3) and (4), below.

To summarize, we wish to maximize a welfare function, W , in which the individuals' utility functions enter as arguments, subject to a constraint, F , which is a product-transformation function. In symbols:

$$\text{Max}_{C^1, C^0, x} W(U^1, U^1, \dots, U^1, U^0, U^0, \dots, U^0),$$

where $U^1 = U(C^1, x)$, $U^0 = U(C^0, 0)$, and subject to:

$$F(C, x) = 0,$$

where $C = \sum C^1 + \sum C^0$.

As before, x stands for production (and consumption) of telephone service, and C stands for the quantity produced (and consumed) of all other goods and services.

Suppose, now, that we measure x by the number of telephones installed and serviced. Then, out of the total group membership of h individuals, x individuals are connected by telephone—have access—and the remaining $h - x$ individuals have no access to the system. We can then state the optimization problem as follows:⁵

$$\text{Max}_{C^1, C^0, x} W = W[xU^1, (h - x)U^0] \quad (3)$$

subject to:

$$F[xC^1 + (h - x)C^0, x] = 0. \quad (4)$$

Assuming that an interior solution exists ($1 < x < h$), we can derive the necessary conditions for an optimum through the Lagrangian function:⁶

$$\text{Max}_{x, C^1, C^0, \lambda} L = W - \lambda F \quad (5)$$

$$\frac{\partial L}{\partial C^1} = W_1 x U_c^1 - \lambda F_c x = 0 \quad (6)$$

$$\frac{\partial L}{\partial C^0} = W_0 (h - x) U_c^0 - \lambda F_c (h - x) = 0 \quad (6a)$$

⁵ Note that in formulating the problem this way, we are using a more specific welfare function, and that some individuals may have differently composed bundles of goods in different allocations.

⁶ The optima will be global under adequate concavity and convexity assumptions on W and F .

$$\frac{\partial L}{\partial x} = W_1(U^1 + xU_x^1) - W_0U^0 - \lambda \frac{dF}{dx} = 0 ; \quad (6b)$$

differentiating with respect to λ ,

$$\frac{\partial L}{\partial \lambda} = F = 0, \quad \text{which is the same as (4).}$$

From (6) and (6a) we find the following *optimality condition*:

$$W_1U_c^1 = W_0U_c^0. \quad (7)$$

In words, the last unit of consumer goods and services gives the same welfare, as measured by the group welfare index, no matter to which individual it is allocated.

Using all three relations (6), (6a), and (6b), we obtain the following condition:

$$\left(\frac{U^1}{U_c^1} - \frac{U^0}{U_c^0} \right) + x \frac{U_x^1}{U_c^1} = (C^1 - C^0) + \frac{F_x}{F_c}. \quad (7a)$$

The interpretation of (7a) goes as follows. The last term on the right-hand side is the usual marginal rate of transformation (MRT); in our case, it measures the marginal quantity of C (goods and services) which must be given up, in order for the economy to produce (and service) one more telephone. The last term on the left-hand side states the marginal rate of substitution between telephone service and other goods and services (MRS). As is seen, this MRS is multiplied by x , the number of individuals who have access to the telephone system. Thus, the interpretation to be given to this term is that of a public good. For the telephone subscriber, access to the telephone system is a public good, plain and simple.

What shall we make of the two bracketed terms, one on the left-hand side and one on the right-hand side? Although the formulation of the problem that we have chosen enables us to use calculus in the optimization procedure, and thus to handle the discreteness, or (0, 1) property of access/no access, the discrete nature of the problem does show up in these two terms. To begin with the bracketed term on the right-hand side, as a marginal individual joins the club—enters the telephone system—he will give up a quantity C^0 and gain a quantity C^1 . The difference between the two, if positive, must be produced (and if negative helps to reduce the marginal rate of transformation). Analogously, the bracketed term on the left-hand side shows the net difference in total welfare obtained by having the marginal individual switch from the basket $(C^0, 0)$ to the basket (C^1, x) . Since all terms in this condition are expressed in units of C , both U^1 and U^0 in this bracketed term are weighted by the respective marginal utilities of C .

□ **Introducing transaction costs.** In the last few years researchers have directed growing attention to theories of the allocation of time, the interest having been sparked perhaps especially by the work of Gary Becker.⁷ Increasing attention, although generally unrelated to the first-mentioned strand of ideas, has also been given to the concept of

⁷ See [3]. Compare also [2].

transaction cost, in particular as related to market failure.⁸ In this paper, we shall relate some of these ideas in a very simple manner to the telephone system. Our procedure will be to extend the model discussed above by introducing time-budget constraints.

Assume that every individual—no matter whether he has a telephone or not—spends an average of τ hours (per unit time period) walking, or driving, to each individual who cannot be reached by telephone. Thus, each individual in G_0 , the group of individuals without access to telephone, spends $(h - 1)\tau$ hours, on the road, in order to have face-to-face conversation with other individuals. Similarly, the x individuals in group G_1 , who do have access to the telephone system, must each spend a total of $(h - x)\tau$ hours getting in touch with the individuals who do not have telephones. To begin with, we shall assume that all other time that an individual has available (working time being fixed) can be enjoyed—for instance, in conversation with other individuals over the telephone or face-to-face—and therefore enters the utility function.

Thus, to our previous formulation, we now add the following two time-budget constraints:

$$\bar{t} - t^0 - (h - 1)\tau = 0, \text{ for all individuals in } G_0 \quad (8)$$

$$\bar{t} - t^1 - (h - x)\tau = 0, \text{ for all individuals in } G_1. \quad (9)$$

Here, $\bar{t}$ denotes the total time available, per individual, and t^i , $i = 0, 1$, denotes the enjoyable time per individual respectively, without or with access to the telephone system.

As before, we set out to maximize $W = W[xU^1, (h - x)U^0]$, where we now have:

$$U^1 = U(C^1, x, t^1) \text{ for all individuals in } G_1$$

and

$$U^0 = U(C^0, 0, t^0) \text{ for all individuals in } G_0$$

and where the optimization is now subject to three constraints, namely (4), (8), and (9). The Lagrangian becomes:

$$\begin{aligned} \text{Max}_{\substack{C^0, C^1, t^0, t^1, x \\ \lambda, \sigma^0, \sigma^1}} L = & W[xU^1, (h - x)U^0] - \lambda F - \sigma^0[\bar{t} - t^0 - (h - 1)\tau] \\ & - \sigma^1[\bar{t} - t^1 - (h - x)\tau]. \end{aligned} \quad (10)$$

From the necessary conditions for an optimum, we once again obtain the optimality condition (7), above. However, the condition (7a) will now have one more term:

$$\left(\frac{U^1}{U_e^1} - \frac{U^0}{U_e^0} \right) + x \frac{U_x^1}{U_e^1} + \tau x \frac{U_t^1}{U_e^1} = (C^1 - C^0) + \frac{F_x}{F_e}. \quad (11)$$

Here, the last term on the left-hand side is new, in comparison with the previously obtained condition (7a). The interpretation of this new term can be made as follows. As the marginal telephone is connected to the system, each individual in G_1 saves time by not having to walk or drive, any longer, to the new telephone subscriber. Since there are x members in G_1 , the total time-saving is found by summation over all x , and the thus gained time is evaluated by the marginal rate of substitution U_t^1/U_e^1 . In this way, *two* terms in the condition

⁸ See [1], [5], and [11].

(11) bring out the public-good property of telephone service, both terms being *summations* over marginal rates of substitution. One term relates to (potential) access and the other to (effective) savings in transaction costs.

One more extension of the model can now easily be made, namely through the explicit consideration of working time as a variable. Assuming that everyone of the x individuals in G_1 (with access to the telephone system) spends T^1 hours at work, and that each one of the $(h - x)$ individuals in G_0 spends T^0 hours at work, we have the definitional relation:

$$T = xT^1 + (h - x)T^0, \quad (12)$$

where T denotes the total amount of time spent at work by all the h individuals.

Through appropriate insertions of these variables into the transformation function F and into the two time-budget constraints, we can again set out to maximize welfare, and this time (as usual assuming that a solution exists for $1 < x < h$) we find two optimality conditions:

$$W_1 U_c^1 = W_0 U_c^0 \quad (7)$$

and

$$W_1 U_t^1 = W_0 U_t^0. \quad (13)$$

From these we derive a standard efficiency result for private goods:

$$\frac{U_c^1}{U_t^1} = \frac{U_c^0}{U_t^0}. \quad (14)$$

In the condition corresponding to (11), we obtain:

$$\begin{aligned} \left(\frac{U^1}{U_c^1} - \frac{U^0}{U_c^0} \right) + x \frac{U_x^1}{U_c^1} + x\tau \frac{U_t^1}{U_c^1} \\ = (C^1 - C^0) + \frac{F_x}{F_c} + (T^1 - T^0) \frac{F_T}{F_c}. \end{aligned} \quad (15)$$

The only difference between (15) and (11) is the last term appearing on the right-hand side in (15), $(T^1 - T^0)F_T/F_c$. This term can be interpreted as a marginal shift in the transformation curve resulting from the change in the working time of the marginal individual who becomes a telephone subscriber.

From the necessary conditions for equilibrium in this scenario, we can also derive the following relationship:

$$-\frac{F_T}{F_c} = \frac{U_t^1}{U_c^1}. \quad (16)$$

In (15), the savings in transaction costs—indicated by the last term on the left-hand side—show up as a welfare gain accruing to individuals in the subset G_1 . Alternatively, using (16) to substitute for U_t^1/U_c^1 in (15), one can have these time-savings converted into more production of C .

To summarize, by using

- (1) the production possibility frontier, F ;
- (2) the two time-budget constraints, for each individual in G_0 and G_1 , respectively; and
- (3) the four relations (7), (13), (15), and (16),

we have a system of seven independent equations. With the appropriate curvature assumptions (that the production possibility set is convex and that the welfare objective function is concave), these seven equations provide us with a solution in the seven variables C^1 , C^0 , t^1 , t^0 , T^1 , T^0 , and x .

Thus, our scheme provides the omniscient planner with a calculation to confirm what he or she presumably already knows, namely the optimal number of telephones in the system. This optimal point is also an efficient point. More important in the context of the present paper, our findings reveal that having access to the telephone system is a public good, and further that new entrants to the system provide "gains" to the system (in the form of savings in transaction costs), which are similarly public-good flavored.

□ **Pricing.** To bring out the essence of the static resource allocation problem involving telephone service, we have here proceeded along Samuelsonian lines. The necessary conditions for equilibrium have been derived without reference to any particular form of economic organization, such as perfect competition. There are no prices or incomes appearing (explicitly) in our problem formulation. It is of course a question of great interest and significance whether some such "signalling system" as contained in perfect competition or in (decentralized) perfect planning would sustain a configuration of resource allocation comparable to the results above—comparable in the sense of also being efficient.

In order to investigate this problem, one must consider in some detail and specification the optimizing behavior of micro-units, firms, and households. Assuming that all individuals have identical utility functions, possess identical endowments, and are faced by the same prices, our conjecture is that a pseudo-*laissez-faire* framework (to use Samuelson's words) would lead to an allocation of $x = 0$ or $x = h$, that is, where either nobody has a telephone or everybody has. Then, at least differentiation of prices and endowments would be required to sustain a pseudo-equilibrium with an interior solution in x .

Considering the production and supply side of the economy, even if we were to assume that the individual firms respond as profit-maximizers to parametric prices, there would still be a difficulty to handle the pricing related to the discrete terms referring to access.

If we were to have individual pricing and lump-sum payments varying in size (and sign) between individuals, such that the income distribution *ex post* would be perfectly compatible with any societal welfare function, the necessary conditions that we would obtain from the profit and utility maximization for the micro-units would still not imply an allocation of resources satisfying the results obtained in the previous section of this paper. It would take a more complicated pricing scheme than that contained in perfect competition to sustain the efficiency conditions. As we saw in a previous section, these conditions would have three components, difficult to handle from a pricing point of view. These three components are:

- (1) a "public good" component (access to the system)
- (2) a "transaction cost" component (involving savings in connection costs), and
- (3) a "discreteness" component.

We have attempted here to analyze some properties of the telephone system, as they relate to the pure theory of public goods. In the process, we have used some highly simplifying assumptions. Although these have enabled us to cast the problems in a general equilibrium framework, the results have come at a price. The analysis has been static. We have been forced to assume that the individual consumer obtains or receives *either* the bundle (C^1, x, t^1) or the bundle $(C^0, 0, t^0)$.

In the next section, we shall attempt to remedy these deficiencies. We shall consider an important *dynamic* aspect of public goods, and we shall introduce in a meaningful way *incomes* and the *distribution* of incomes. In the process, we shall redirect our attention away from the production side and towards the demand side.

■ We shall now construct a growth model of the demand for telephones. We employ a particular public-good property of the telephone system, together with a set of quite reasonable and not unduly restrictive assumptions, in order to derive a self-sustaining growth process. Time will be treated discretely.

The public-good property which we shall employ can more appropriately be displayed in a dynamic framework. Suppose that over a time interval, dt , a number of new subscribers, say, $\dot{x}dt$, obtain telephones. Then, in the terms of our previous analysis, the associated argument that enters their preference functions is $(x + \dot{x}dt)$. In other words, potential use of the telephone system—access, in our terminology—is cumulative, like a stock, and is not in the nature of a flow. Thus, and disregarding for simplicity any “decay” process involved, we can regard our access variable, x , as related to time in the following way:

$$x(t) = \int_0^t \dot{x}(s)ds .$$

As before, let h denote the size of the total population, and p_x denote the subscription price of telephone service. Let y denote the level of an individual’s income and $\Psi(y)$ be a continuous, strictly increasing cumulative distribution of income, expressing the share of the total population with income below y .

(A1) Our first assumption, then, is that the total population h , the distribution of income $\Psi(y)$, and the price of telephone subscription p_x are all independent of time, and thus *stationary*.

We realize that by assuming a growing population, growing incomes, and possibly a changing income distribution and changing relative prices, our analysis would become more “realistic.” But that is not our purpose. Instead, we wish to show that *even* if there is no population and income growth, etc., we can still obtain a sustained growth process.

(A2) Our second assumption is that the distribution of telephones among individuals is such that the order of new subscribers entering the telephone system is the order of their income:

$$x(t) = h[1 - \Psi z(t)] . \quad (21)$$

Here, $x(t)$ is the total number of telephones in period t , and $z(t)$ is

3. Dynamics of the demand for telephones

the income level *beyond* which all individuals have access to the telephone system in period t .

(A3) Our third assumption is that the preferences of individuals can be represented by a utility function $U(c, \delta x)$, continuous and strictly increasing in c and in δx , and where c is the consumption of private goods by the individual, x is the number of telephones to which the individual has access, and

$$\delta = \begin{cases} 1 & \text{if the individual has } a \\ 0 & \text{no} \end{cases} \text{ telephone.}$$

(A4) Our fourth assumption is that at any time each individual maximizes his utility subject to his budget constraint:

$$\text{Max}_{c(t+1), \delta} U[c(t+1), \delta x(t)] \text{ subject to } c(t+1) + \delta p_x = y(t+1).$$

Here, and without any loss of generality, the price of the private goods has been set at unity.

(A5) Finally, we assume that $x(0)$ and $x(1)$ are given, such that $x(1) > x(0)$. In other words, we assume that some factor exogenous to the model has initiated a change in the number of telephones connected.

Proposition. Given (A1)–(A5), the sequence of $x(t)$ increases.

To prove this proposition, we reason by recurrence. Assume $x(t) > x(t-1)$, then $U[z(t) - p_x, x(t)] > U[z(t) - p_x, x(t-1)]$. This follows from the assumption that U is strictly increasing in x .

By continuity of U , there exists $y < z(t)$, such that $U[y - p_x, x(t)] > U[z(t) - p_x, x(t-1)] = U[z(t), 0]$. Here, the last equality is obtained by the definition of $z(t)$. Since $y < z(t)$, it follows from the assumption that U is strictly increasing in c , that $U(y, 0) < U[z(t), 0]$.

Thus, there exists $y < z(t)$ such that $U[y - p_x, x(t)] > U(y, 0)$. This means that in period $(t+1)$ new demand is generated for telephone service from individuals with incomes somewhat below $z(t)$. Accordingly, $z(t+1) < z(t)$, and since $\Psi(y)$ is strictly increasing, it follows that $x(t+1) > x(t)$.

To summarize, it has been assumed that $x(1) > x(0)$, and it has been proven that if $x(t) > x(t-1)$, then $x(t+1) > x(t)$. Thus, the sequence in x is increasing. Being bounded above by h , the sequence converges. It converges either to h or to some x^* , $x^* < h$, which, if it exists, must satisfy:

$$U\left[\Psi^{-1}\left(1 - \frac{x^*}{h}\right) - p_x, x^*\right] = U\left[\Psi^{-1}\left(1 - \frac{x^*}{h}\right), 0\right].$$

Obviously the shape of the functions U and Ψ will influence the value to which the sequence converges.

To conclude this section, let us consider an illustration of the theoretical results, through the use of *specific* functions U and $\Psi(y)$. We were intrigued by the fact that the “S”-shaped logistic curve had been applied in some empirical studies to model the growth of the Bell System residence telephones.⁹ Except for the concept of a “saturation level,” and the use of “contagious disease” analogies, there did not seem to be much theoretical justification for the choice of the logistic curve.

⁹ See [4].

So, using our dynamic public-good theory as an anchor, we set out to find a utility function and an income distribution function which would, on the one hand, satisfy the standard economic theory assumptions imposed on such functions and, on the other hand, generate a logistic curve, when inserted in our analytical framework.

Consider, then, the following cumulative distribution-of-income function,¹⁰

$$\Psi(y) = e^{-b/y}, b > 0$$

and the following utility function,

$$U(c, \delta x) = U_1(c) + U_2(\delta x),$$

where

$$U_1'(c) = \frac{h}{p_x} (1 - e^{-b/c})$$

$$U_2(\delta x) = \delta[(1 + ah)x - ax^2], \quad a > 0 \quad a < \frac{1}{h}.$$

Here, both U_1 and U_2 are strictly increasing and concave. Then, it can be shown¹¹ that an accumulation of telephones will occur, governed by the differential equation:

$$\frac{\dot{x}}{x} = a(h - x).$$

The solution of this differential equation yields a logistic curve.

Our result shows that a growth process in the demand for telephones can be explained jointly by the distribution of incomes and the cumulative public-good property inherent in the telephone system, which we have described and analyzed. Recalling our assumptions of a constant population, constant incomes, a stationary income dis-

¹⁰ The corresponding density function is skewed in the (empirically) appropriate way.

¹¹ Using a continuous approach, the fundamental equation defining the process is

$$U[z(t + dt) - p_x, x(t)] = U[z(t + dt), 0],$$

and can be rewritten as

$$U_2(x) = U_1 \left[\frac{-b}{\log \left(1 - \frac{x + dx}{h} \right)} \right] - U_1 \left[\frac{-b}{\log \left(1 - \frac{x + dx}{h} \right)} - p_x \right],$$

or for incomes sufficiently larger than p_x ,

$$U_2(x) = p_x U_1' \left[\frac{-b}{\log \left(1 - \frac{x + dx}{h} \right)} \right]$$

or

$$(1 + ah)x - ax^2 = x + dx,$$

leading to the differential equation

$$\frac{\dot{x}}{x} = a(h - x),$$

integrated in $x(t) = \frac{h}{1 + \frac{h + x(0)}{x(0)} e^{-hat}}$

The behavior of $x(t)$ is thus the consequence of the assumptions (A1) to (A5) and of the specific shape of the functions U and ψ .

tribution, and no change in relative prices, we venture the observation that the result is interesting. The use of a specific utility function and a specific income distribution are illustrative: from them we derive the logistic curve which has been used in descriptive studies of the demand for residence telephones.

4. Concluding remarks

■ It goes without saying that only a few of the many interesting analytical aspects of the telephone system have been analyzed in this paper. The pricing problems deserve much closer scrutiny than we have afforded them. Peak-load and waiting-line type of problems deserve attention, also, in the public-good context. The point should already be clear from the paper that the welfare gains to society arising from the telephone system—in almost whatever measure we care to express them—are vastly greater than the costs. A more complicated but related issue, not dealt with in our paper, concerns the rates of growth in benefits and costs over time, and this topic deserves further analysis.

On the other hand, we believe that the aspects which we have been able to analyze pervade to a much broader spectrum of societal phenomena than the telephone system. Many economic phenomena in modern life have the peculiar access/no access property that we have described and analyzed. Other communication systems, as well as various transportation systems, appear to have similar characteristics. The same is true, much more broadly, of what we may loosely call “urban economic systems.” In that context, the uneven distribution of access vs. no access in a population, raises fundamental issues of equity.

References

1. ARROW, K. “Political and Economic Evaluation of Social Effects and Externalities” in J. Margolis, ed., *The Analysis of Public Output*, New York: National Bureau of Economic Research, 1970.
2. ARTLE, R. *Staden som kollektivfenomen*. Stockholm: Allmanna Forlaget, 1972.
3. BECKER, G. S. “A Theory of the Allocation of Time.” *Economic Journal*, Vol. 75, No. 298 (September 1965), pp. 493–517.
4. CHADDHA, R. AND CHITGOPEKAR, S. “A Generalization of the Logistic Curves and Long-Range Forecasts (1966–1991) of Residence Telephones.” *The Bell Journal of Economics and Management Science*, Vol. 2, No. 2 (Autumn 1971), pp. 542–560.
5. HAHN, F. “Equilibrium with Transaction Costs.” *Econometrica*, Vol. 39, No. 3 (May 1971), pp. 417–439.
6. HARSANYI, J. C. “Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility.” *Journal of Political Economy*, Vol. 63, No. 4 (August 1955), pp. 309–321.
7. MILLERON, J. C. “Recent Developments in the Theory of Public Economics.” Working Paper, Center for Research in Management Science, University of California, Berkeley, 1971.
8. SAMUELSON, P. A. “The Pure Theory of Public Expenditure.” *Review of Economics and Statistics*, Vol. 36, No. 4 (November 1954), pp. 387–389.
9. ———. “Pure Theory of Public Expenditure and Taxation” in J. Margolis and H. Guitton, eds., *Public Economics*, London: Macmillan & Co., 1969.
10. SOLTOW, L., ed. *Six Papers on the Size Distribution of Wealth and Income*. New York: National Bureau of Economic Research, 1969.
11. VARAIYA, P. AND ARTLE, R. “Locational Implications of Transaction Costs.” *Swedish Journal of Economics*, Vol. 74, No. 1 (March 1972), pp. 174–183.

A decision analysis with multiple objectives: the Mexico City Airport

Ralph L. Keeney

Assistant Professor of Operations Research and Management
Sloan School of Management
Massachusetts Institute of Technology

This paper reports an analysis done for the Secretaria de Obras Publicas (Ministry of Public Works) of Mexico to help select the most "effective" strategy for developing the airport facilities of the Mexico City metropolitan area to insure quality air service for the remainder of the century. Effectiveness is a complex function including attributes of cost, safety, capacity of the airport facilities, noise levels, social disruption, and access times. A decision analytic model was used for evaluating strategies. The attributes were adapted to account for impacts over time, and probability density functions and a utility function were assessed over the six attributes. Details of these assessments are given. The results and implications of the analyses are discussed. A unique aspect of the study involved the assessment of a multiattribute utility function and its use to aid an important policy decision.

■ Two previous studies of the development of airport facilities of Mexico City had recommended very different alternatives. One¹ concluded that the current airport at Texcoco, five miles from the city center, should be greatly expanded, whereas the other² suggested moving all aircraft operations to a new airport to be built at Zumpango, 25 miles north of the city. Our initial charter was to evaluate the various alternatives, in light of this discrepancy, and to recommend the most effective program for airport development. The alternatives and measures of effectiveness were quite firmly specified by our clients, the Secretaria de Obras Publicas (SOP).³

□ **The alternatives.** The alternatives specified what types of aircraft would operate at which of the two possible sites over the rest of the century. Because of similarities in operating characteristics and functions, the SOP had categorized aircraft as follows: international (I), domestic (D), general (G), and military (M). It was assumed at any one time, each category of aircraft could operate at either site.

This work was done under the auspices of the Secretaria de Obras Publicas and directed by F. J. Jauffred, Director of the Center for Computation and Statistics, and F. Dovali, Head of the Department of Airports. Richard de Neufville of M.I.T., Howard Raiffa of Harvard University, and I were consultants assisting the SOP on the project. Craig Kirkwood and B. Katherine Swartz made many helpful comments on the content of this paper. The total time spent by the consultants on the project was 50 man-days. The work for this article was supported in part by the Office of Naval Research under Contract N 00014-67-A-0204-0056.

¹ In [3].

² In [9] and [11].

³ The SOP is the Mexican Ministry of Public Works.

1. The problem

To account for changes in operating arrangements over the 30-year horizon while keeping the problem manageable, we decided to focus on the three years, 1975, 1985, and 1995, in which changes in the classes of aircraft operating at a site could occur. Thus, an alternative might be "develop the Zumpango site and move general aircraft to it in 1975, shift international to Zumpango in 1985, and operate all classes of aircraft at Zumpango by 1995."

Notice that this gives 16^3 or 4096 alternatives. However, many of these are very similar in nature since, for instance, military operations account for less than five percent of the aircraft volume. Other alternatives as defined above were unreasonable. One would not move all operations from Texcoco to Zumpango in 1975 and back again in 1985, for example. In the final analysis, the total number of alternatives evaluated was approximately 100.

□ **Measures of effectiveness.** To evaluate the alternatives, one needs to specify some measures of effectiveness which explicitly describe possible impacts on each of the important groups concerned with the problem. These groups are the government as builder and operator of the airports, the users of the air facilities, and the nonusers. After lengthy discussions, the following six measures were selected by the SOP for the evaluating effectiveness:

- X_1 $\equiv$ total cost in millions of pesos;
- X_2 $\equiv$ the practical capacity in terms of the number of aircraft operations per hour;
- X_3 $\equiv$ access time to and from the airport in minutes, weighted by the number of travelers from each zone in Mexico;
- X_4 $\equiv$ number of people (including nonpassengers) seriously injured or killed per aircraft accident;
- X_5 $\equiv$ number of people displaced by airport development; and
- X_6 $\equiv$ number of people subjected to a high noise level, in this case 90 CNR⁴ or more.

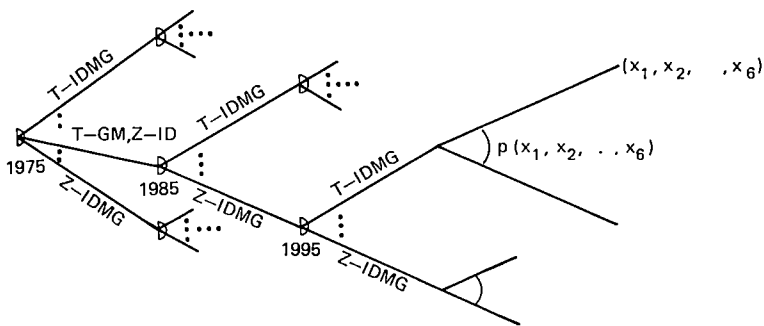
Although there is much overlap, the first two measures account for government's interest, the third for users, and the last three for nonusers. Clearly, these six measures of effectiveness are neither unique nor completely comprehensive. For instance, air pollution considerations are absent. However, the SOP felt the list did include all the important factors for evaluating effectiveness of the proposed alternatives.

2. The basic decision model

■ The basic model is illustrated by the decision tree in Figure 1. An alternative is first specified by defining what classes of aircraft will operate at which site in each of the three years. As a result of the alternative chosen and the events which occur (e.g., demand changes), a consequence $(x_1, x_2, \dots, x_6)$ will eventually result. However, at the time the decisions must be made, one cannot be sure what consequence will result. This uncertainty is quantified with a probability density function, $p(x_1, x_2, \dots, x_6)$, describing the likelihood of each consequence conditional on the alternative chosen.

⁴ The Composite Noise Rating, CNR, is a standard index of noise which combines decibel level and frequency of occurrence. The 90 level was selected by the SOP Department of Airports.

FIGURE 1
THE BASIC DECISION MODEL*



* THE NOTATION T-GM,Z-ID MEANS OPERATE GENERAL AND MILITARY AIRCRAFT AT TEXCOCO AND INTERNATIONAL AND DOMESTIC AIRCRAFT AT ZUMPANGO

There are two important points to note about the problem. First, the alternatives are master plans; they are not adapted to account for critical events (e.g., demand changes, technological developments, etc.) which might occur in the 30-year period. Clearly, such considerations are important. The results of the analysis reported here, both indicate the effectiveness of various proposed airport developments for Mexico City and are a component part of an overall analysis of the previously reported 1971 options open to the Government of Mexico.⁵ This analysis accounted for potential political effects and allowed flexibility to adapt strategies for possible future events.

The second point is that each of the measures of effectiveness must account for the impact over the 30-year period from now to 2000. This required that the consequences $(x_1, x_2, \dots, x_6)$ be adapted to incorporate time effects.

□ **Incorporating time effects.** The costs considered in the model included building and maintenance, but excluded operating costs, since it was felt that these would be approximately the same for any alternative. For each alternative, the present value of these costs was taken as the time dependent attribute of importance. The discount rate used was 12 percent.

For noise, the average number of people annually subjected to aircraft noise levels above 90 CNR was used as the measure of effectiveness. This assumes that it is as undesirable to have one person subjected to these noise levels for two years as to have two different people subjected for one year each. Furthermore it assumes that the undesirability of a certain noise level is the same in any year.

As previously mentioned, safety is measured in terms of the number of people killed or seriously injured per air crash. To adapt this, we chose the average number of people killed or seriously injured per crash over the 30-year time period.

For access time we used the average of the possible access times in the various years weighted by the expected number of users in those years. This assumes each trip to or from the airport by any individual in any year is as important as any other such trip, and that one's preferences for the various access times are stationary over time.

⁵ In [1].

AIRPORT DECISION
ANALYSIS / 103

We reasoned that forcing an individual to relocate because of airport development would be equally undesirable in any year. Therefore, we chose the total number of such people displaced to be the measure of social disruption for the analysis.

Capacity (maximum possible operations/hour) could not be aggregated in any reasonable way to combine impacts in the different years. This is mainly because the relative desirability of various levels of capacity would be very different in different years, since demand would probably be larger in later years. Increasing capacity from 80 to 100 in 1975 may be worth very little, as the additional capacity would rarely be needed. However, this same change in 1995 could be extremely important. Thus, in the 30-year model separate measures of effectiveness for the capacity of 1975, 1985, and 1995 were included.

□ **Outline of the paper.** The impact of the alternatives is discussed in Section 3. For each of the alternatives, we assess a joint probability density function over the measures of effectiveness, discuss the subjective process by which these were assessed, and describe experiments used to gain useful insights.

In Section 4, the multiattributed utility function over the measures of effectiveness is given. This utility function indicates the tradeoffs among the various effectiveness measures. Coupled with the probability density function, it permits one to evaluate the expected utility of each alternative. This “expected utility” is a number which measures the overall “effectiveness” (as it is defined) of each alternative.

The results and conclusions of the analysis are presented in Sections 5 and 6.

All the judgmental assessments concerning probabilistic impact and preferences were provided by members of the SOP. They included the Director of Airports, who is responsible for building and maintaining all the airports in the country of Mexico, the Director of the Center for Computation and Statistics, and the members of their staffs. The assessments were made in group sessions, where a consensus was easily reached because of small original differences.

3. Assessing the probability of various consequences

■ For each of the 16 arrangements for aircraft operation at the two possible sites in 1975, 1985, and 1995, general construction plans indicated where runways, support facilities, and access facilities would have to be built. These plans translated the feasible alternatives specified in Section 2 into designs meaningful to airport planners and government officials.

To gain insight before making these assessments, the SOP conducted various experiments. One, designed to gather data on access times involved: (1) dividing Mexico City into approximately ten zones; (2) specifying the airport usage characteristics of the inhabitants of each region; and (3) driving to the two airport sites from each zone in different weather conditions at different times of day. These data on travel times and usage characteristics provided the information necessary to assess reasonable probability density functions for access times under the various alternatives.

In a similar way, detectors were located at various spots in the city to determine the noise levels caused by aircraft. By analyzing current and projected flight paths, superimposed over aerial photos

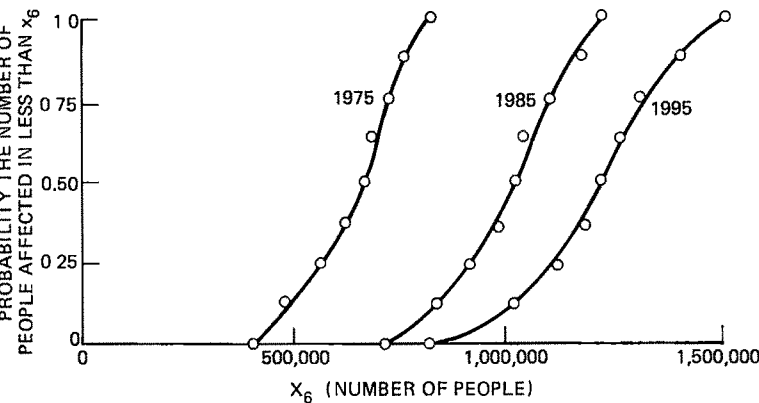
of the city, and the population densities of the affected areas, the SOP acquired a good indication of the noise impacts of various alternatives. These helped to assess probability densities for the number of people subjected to a specified noise level.

By superimposing the various plans for construction over aerial views of the city, the SOP could easily identify the areas in which, people would have to be relocated, given adoption of a particular alternative. The population of those areas was accounted for, and this provided information for assessing the number of people who would be displaced.

On the other attributes—cost, capacity, and safety—the assessments relied more on the expert judgment of the Director of Airports. Of course, the construction plans and, in the case of safety, population densities in the various flight paths were also useful.

□ **One-year assessments.** The probability density functions were assessed using the fractile method described by Raiffa.⁶ Figure 2 illustrates the method by example. Consider the possible 1975 noise

FIGURE 2
NUMBER OF PEOPLE SUBJECTED TO NOISE LEVELS OVER 100 CNR
WITH THE "ALL TEXCOCO" OPTION



impact of the operating arrangement of "all classes of aircraft at Texcoco." First, the maximum and minimum number of people subjected to 90 CNR or greater was specified as 800,000 and 400,000, respectively. Next the 0.5 fractile was evaluated as 640,000. In the judgment of the SOP, this meant that the likelihood is one-half that the number of people impacted by 90+ CNR (denoted by X_6^{75}) would be less than 640,000. The interval between 400,000 and 640,000 was then divided into equally likely parts by choosing the 0.25 fractile as 540,000. The 0.75 fractile was 700,000. Finally, each of the quartiles was divided into equally likely parts in a similar manner. The dots on Figure 2 indicate the assessed fractiles, and the smoothed line represents the probability density function describing the possible impact of noise for the "all Texcoco" option. As illustrated, for any given year the probability of the impact being between any two adjacent fractile points should be the same, that is 0.125. Thus, to check consistency of the assessments, we asked the SOP if in fact their judgmental probabilities of falling into any of the eight such ranges of impact were the same. The SOP adjusted their assess-

⁶ In [6].

ments until no more discrepancies could be found. Figure 2 indicates the final adjusted curves.

□ **The 30-year assessments.** By aggregating the three yearly assessed impacts for each measure of effectiveness (except capacity) in the manner indicated in Section 2, we could calculate the probability density functions over the measures to account for impact over time. For instance, with noise, if we define

$$X_6 \equiv \frac{X_6^{75} + X_6^{85} + X_6^{95}}{3},$$

where X_6^i is the number of people subjected to noise levels over 90 CNR in year i , then by using the probability density functions assessed for the X_6^i for a particular strategy, it is straightforward to derive the probability density function for X_6 . This represents what we have taken to be the overall impact of that particular strategy in terms of noise.

□ **Probabilistic independence assumptions.** In conducting the assessments over one attribute at a time, we were explicitly assuming that conditional on each alternative, the six attributes were probabilistically independent. For some of the attributes, this assumption seems appropriate. For instance, conditional on any given alternative, noise and access time considerations are probably independent of the other attributes. On the other hand, safety considerations may depend on capacity. The lower the capacity, the more often the airport will be operating under hazardous conditions.

The more important assumption with regard to these assessments was that impacts in separate years were probabilistically independent, conditional on the given alternative. This is clearly not true. For instance, for the "all Texcoco" alternative, if we found that 800,000 rather than 400,000 people were subjected to high noise levels in 1975, we would likely feel that more people would be affected by noise in 1985.

The main reason for these independence assumptions is simplicity. The justification is that, based on sensitivity analyses, the same types of strategies appear to be most effective over wide ranges of probabilistic impacts. If this had not been the case, we might have been forced to build a sophisticated simulation model of aircraft operations within the Mexico City area which would generate more realistic joint probability density functions.

One could argue that given the oversimplifying probabilistic assumptions and the insensitivities, it might have been just as accurate and more simple to use point estimates of the impacts rather than probability densities. In retrospect, this seems quite reasonable. However, one does not avoid any of the assumptions made in our analysis, and in addition, no account is made for the possible distribution of impact for the single attributes. Our approach forces an explicit recognition of this distribution by the decision makers. Also, before the analyses, the lack of sensitivity of the types of effective strategies to the attribute values was not known. A sensitivity analysis using point estimates could have indicated this, however. The strongest reason for maintaining the detail of using probability density func-

tions was that the SOP wanted to avoid potential criticism of the analysis due to exclusion of the distribution of effects.

■ Once we had probability density functions which adequately described the impact of alternate strategies in terms of our six measures of effectiveness, the next step was to assess a utility function $u(x_1, x_2, \dots, x_6) \equiv u(\underline{x})$ over these measures. Given the probability densities describing the impacts of an alternative and given the utility function, one can calculate the expected utility for that alternative. This expected utility is then the measure of desirability of the alternative, and the decision maker should prefer the alternative with the greatest expected utility. This result follows directly from the "axioms of rational behavior" specified by von Neumann and Morgenstern.⁷

Experience has indicated that utility assessments varying more than two attributes simultaneously are extremely difficult for decision makers. Hence, our general procedure for assessing the six attribute utility function is as follows. First we identify assumptions which allow us to find a function f such that

$$u(\underline{x}) = f[u_1(x_1), u_2(x_2), \dots, u_6(x_6)], \quad (1)$$

where $u_i(x_i)$ is a utility function over attribute X_i . Then to acquire $u(\underline{x})$, we must assess the six $u_i(x_i)$ and scaling factors, which indicate the tradeoffs among attributes. This is much easier than any method to assess $u(\underline{x})$ directly. Of course, one of the first steps in the assessment procedure must include a verification of the assumptions which led to (1).

In the next subsection we describe the assumptions used to obtain a utility function of the form (1) for the Mexico City Airport problem. Then we indicate how these assumptions were verified. Details on the assessment of the single attribute utility functions and scaling factors are also given.

□ **The assumptions.** The two important assumptions which we made and later verified dealt with the concepts of preferential independence and utility independence.⁸ Designate $X \equiv X_1 \times X_2 \times \dots \times X_6$, $X_{ij} \equiv X_i \times X_j$, and X_{356} , named "ex not three five," as $X_1 \times X_2 \times X_4 \times X_6$, for example. A small x_{356} will represent a member of X_{356} and x_i a member of X_i .

We will say X_i is preferentially independent of X_{ij}^- if when (x_i, x_j, x_{ij}^-) is preferred to (x_i', x_j', x_{ij}^-) for some x_{ij}^- , then it must be true for all x_{ij}^- . This implies that $u(x_i, x_j, x_{ij}^-) > u(x_i', x_j', x_{ij}^-)$ for all x_{ij}^- , if it is true for any. If X_i is preferentially independent of X_{ij}^- , then the absolute preferences (i.e., rankings) of pairs of consequences x_i , must be the same for all x_{ij}^- . Furthermore, the tradeoffs between X_i and X_j , and therefore the indifference curves over them, will not depend on the amount of x_{ij}^- . The concept of preferential indepen-

4. Assessing the multiattributed utility function

⁷ In [10].

⁸ These are the same concepts referred to as weak conditional utility independence and strong conditional utility independence respectively in Raiffa [7]. Discussions of utility independence and some of its implications are found in Keeney [4, 5].

dence concerns what are sometimes referred to as the ordinal preferences of the decision maker.

Utility independence, on the other hand, concerns the cardinal preferences of the decision maker. Let us define x_i to be a member of X_i which is the Cartesian product of all the attributes except X_i . We will define X_i to be utility independent of X_i if the decision maker's preferences over lotteries involving only x_i , with x_i held fixed, do not depend on the amount of x_i . A fundamental result of utility theory is that utility functions are unique up to positive linear transformations. Thus, when and only when utility functions are positive linear transformations are the decision maker's preferences over lotteries the same. This means that utility independence implies that conditional utility functions over X_i are positive linear transformations of each other, and so we can write

$$u(x_i, x_i) = g(x_i) + h(x_i)u_i(x_i), \quad (2)$$

where g and $h > 0$ are scalar functions.

□ **Verifying the assumptions.** Let us illustrate how we verified the preferential independence assumptions used in our work. As an example, consider whether safety (X_4) and noise (X_6) are preferentially independent of the other attributes. We began by asking questions to find what the amount of safety x_4 was, such that $(x_4; 2500)$ was indifferent to $(1; 1,500,000)$, while the other attributes were fixed at a desirable level. That is, x_4 people seriously injured or killed in an accident and 2,500 people subjected to high noise levels are indifferent to one person seriously injured or killed and 1,500,000 subjected to a high noise level. After "converging," the amount of x_4 was chosen as 300. The exact number is not important for verifying the assumptions, but our interest is in whether it changes as the other four attributes vary. So we next set these four attributes at undesirable levels and asked the same question and again elicited 300 as response.

Then we asked if this would in general be true for any values of the other four attributes, and the response was: "The answer would always be the same, given the other attributes were in a static condition." In fact the respondent stated this would be the case concerning any tradeoffs between safety and noise. Hence, we concluded safety and noise were preferentially independent of the other attributes.

By going through identical procedures, we verified that capacity and cost were preferentially independent of the remaining attributes, as were displacement and access time. By this time, the man answering the questions, an assistant to the Director of Airports, was in a position to state that the tradeoffs between any two attributes did not depend on the amounts of the other attributes. The assumptions were then also verified with other staff members of the SOP, including the Director of Airports.

The same general approach was used in verifying the utility independence assumptions that X_i was utility independent of X_i for all $i = 1, 2, \dots, 6$. As an example, suppose we wanted to find out if access time X_3 was utility independent of X_3 . The other five attributes were set at desirable levels, and the conditional utility function over access time from 12 to 90 minutes (the range originally specified by the SOP) was assessed. We found 62 minutes indifferent to a 50-50 lottery yielding either 12 or 90 minutes. Then we changed

the amounts of the X_3 attributes to less preferred amounts and repeated the question. Again, an access time of 62 minutes was indifferent to a 50-50 lottery yielding either 12 or 90 minutes. A general question indicated this would be true for any fixed amounts of X_3 . We found that relative preferences for any lotteries and consequences involving uncertainties only about access time would not depend on the other attributes. Thus, access time was indeed utility independent of the other five attributes.

This condition was verified for all six attributes with both the Director of Airports and members of his staff. In all of these verification procedures, an attempt was made not to lead the respondent to answers he would not have arrived at otherwise. This appears to have been done successfully. However, since the preferences expressed are subjective and represent the views of the respondents only, there is no wrong or right to them. Since preferences may vary with time, such questioning of the same people may lead to different conclusions at another point in time. The preferences indicated by the individuals questioned do seem to represent their "true" preferences at this time, and hence, the assumptions are appropriate for the problem.

□ **Forms of the utility function.** The main theoretical result used to obtain the utility function is the following:

Theorem. If X_i is preferentially independent of X_j for all i, j and if X_i is utility independent of X_i for all i , then $u(x_1, x_2, \dots, x_6)$ is either an additive or a multiplicative function of the $u_i(x_i)$.

The proof of this result is found in Keeney.⁹

As indicated, the requisite preferential and utility independence assumptions were verified, so we knew that $u(x_1, x_2, \dots, x_6)$ was either additive or multiplicative. The exact form of the utility function u , scaled from zero to one, is

$$\begin{aligned}
 u(x_1, x_2, \dots, x_6) = & \sum_{i=1}^6 k_i u_i(x_i) + K \sum_{i=1}^6 \sum_{j>i}^6 k_j k_i u_i(x_i) u_j(x_j) \\
 & + K^2 \sum_{i=1}^6 \sum_{j>i}^6 \sum_{n>j}^6 k_n k_j k_i u_i(x_i) u_j(x_j) u_n(x_n) \\
 & + \dots + K^5 k_1 k_2 k_3 k_4 k_5 u_1(x_1) u_2(x_2) \dots u_6(x_6), \quad (3)
 \end{aligned}$$

where u_i is a utility function over X_i scaled from zero to one, k_i is a scaling factor for u_i , and K is another scaling constant. The k_i , which must be between zero and one, can be interpreted as the utility u assigned to a consequence with all its attributes except X_i set at their least preferable amount and X_i set at the most preferable amount.

The value of K can be found from the values of the k_i . When $\sum k_i = 1$, then $K = 0$ and (3) reduces to the additive form

$$u(x_1, x_2, \dots, x_6) = \sum_{i=1}^6 k_i u_i(x_i). \quad (4)$$

When $\sum k_i \neq 1$, then $K \neq 0$ so we can multiply each side of (3) by

⁹ In [4]. AIRPORT DECISION
ANALYSIS / 109

K , add one to the results, and factor to get the multiplicative form

$$Ku(x_1, x_2, \dots, x_6) + 1 = \prod_{i=1}^6 [Kk_i u_i(x_i) + 1]. \quad (5)$$

Thus, the additive form can be considered as a limiting case of the multiplicative utility function. A complete discussion of the multiplicative utility function is found in Keeney.¹⁰

□ **Assessing the $u_i(x_i)$'s.** Each of the single attribute utility functions was assessed by the lottery technique discussed by Schlaifer.¹¹ Let us illustrate this technique by assessing preferences for access time.

The first step involved obtaining maximum and minimum values for access time. By questioning the members of the SOP, we found that the range should go from 12 minutes to 90 minutes, where shorter access times were preferable to longer ones. Thus, to keep with our scaling convention where the utility functions ranged from zero to one, we set

$$u_3(90) = 0 \quad (6)$$

and

$$u_3(12) = 1. \quad (7)$$

From questions to check whether X_3 was utility independent of X_3 , recall that we found 62 minutes indifferent to a lottery, which we will denote by [12; 90], yielding either 12 or 90 minutes, each with probability 0.50. Hence, the utility assigned to 62 minutes, the certainty equivalent for the lottery is

$$u_3(62) = 0.50u_3(12) + 0.50u_3(90) = 0.50. \quad (8)$$

Since 62 is greater than the expected access time 51 of the lottery [12; 90], this original assessment indicated that the utility function might exhibit risk aversion. In this context, risk aversion means that the expected amount $\frac{x_3 + x_3'}{2}$ of any lottery $[x_3; x_3']$ would be pre-

ferred to that lottery. By asking about specific lotteries as well as about the general case, we found that the decision makers were risk averse in the attribute access time. This implied that the utility function would necessarily have to be concave as indicated in Figure 3.

By asking more questions to find certainty equivalents of additional lotteries, other points on $u_3(x_3)$ were specified. For instance, we found 40 minutes indifferent to [12; 62] and 78 minutes indifferent to [62; 90], so

$$u_3(40) = 0.50u_3(12) + 0.50u_3(62) = 0.75, \quad (9)$$

and

$$u_3(78) = 0.50u_3(62) + 0.50u_3(90) = 0.25. \quad (10)$$

Then a utility curve was smoothed through the empirically assessed points.

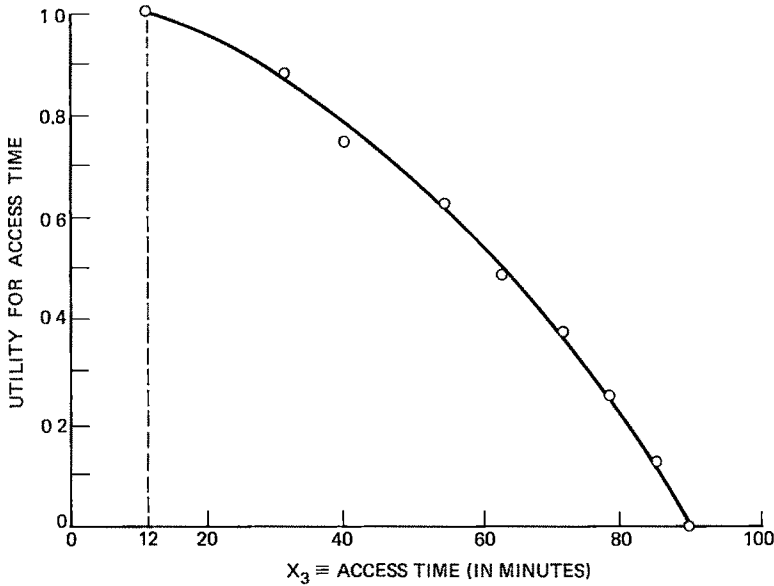
Five points should be sufficient to fair in most utility functions provided it is known what type of risk characteristics the utility function possesses. Such characteristics, along with monotonicity, specify the general form of the curve. These, together with five points on the

¹⁰ In [4].

¹¹ In [8].

FIGURE 3

UTILITY FUNCTION FOR ACCESS TIME



utility function assessed to cover its range, almost completely restrict the shape of the utility function.

However, there is no mathematical test to see if a utility function is “correct.” Since such a function is subjective, we use consistency checks to validate it. Additional questions are asked about the decision maker’s preferences, and his responses are compared with the implications of the utility function. If they are consistent, we develop more confidence in the utility function. If they are inconsistent, the inconsistencies are discussed, and part or all of the assessment is repeated.

Procedures similar to those described above were also used to assess utility functions for cost, safety, displacement, and noise. The results are illustrated in Figure 4. However, as mentioned earlier, no single measure was found to combine capacities in different years. Thus, it was necessary to assess the capacity utility function $u_2(x_2)$ differently.

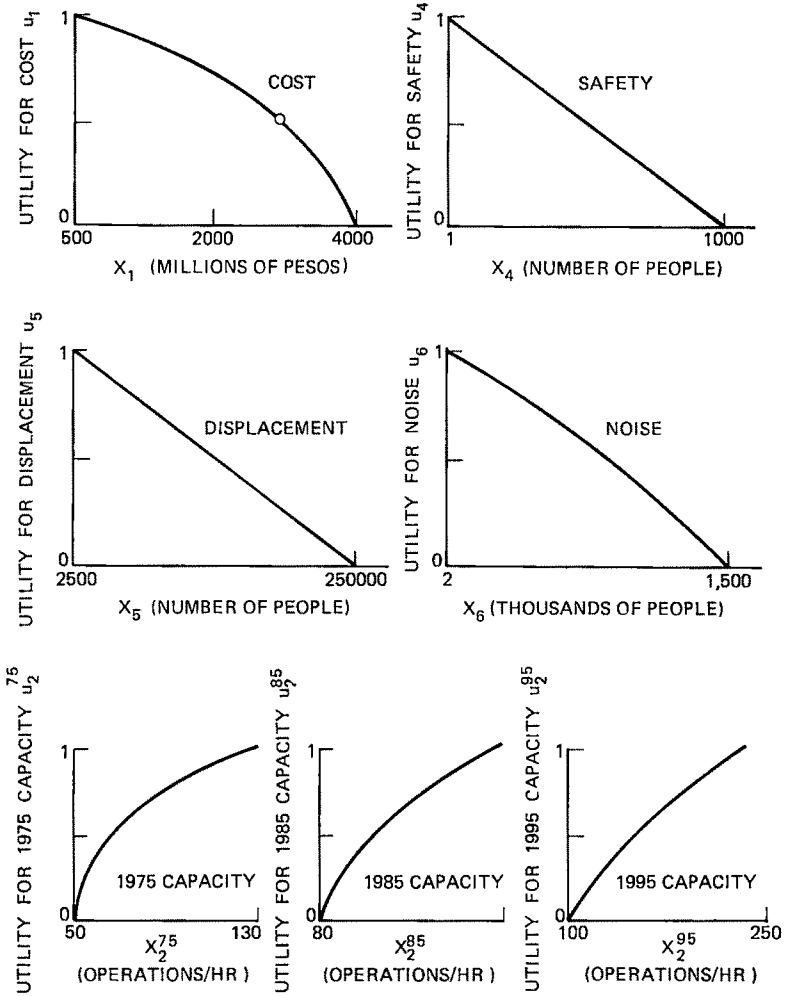
Although the general shapes of the utility functions for access time, cost, and noise seem reasonably consistent with intuition, the linear relationships for safety and displacement may not. For instance, one might expect that the utility function for safety would be risk averse, since governments usually abhor large numbers of deaths resulting from single tragedies due to the political impacts of such tragedies. However, our measure of effectiveness in this problem was not meant to capture these political factors. Roughly speaking, if one says each life is equally important, then alternatives with the same expected number of people killed or seriously injured should be equally undesirable in this respect. This was the attitude taken by the SOP in the assessments, so $u_4(x_4)$ is linear.

□ **The capacity utility function.** Capacity x_2 is a vector $(x_2^{75}, x_2^{85}, x_2^{95})$, where x_2^{75} is the capacity in 1975, etc.

The first step in assessing $u_2(x_2)$ was to identify the minimum and maximum possible airport capacities for each year: 1975, 1985, and

FIGURE 4

THE UTILITY FUNCTIONS



1995. These were 50 and 80, 100 and 130, and 200 and 250, respectively. Clearly more capacity in any given year was preferred to less capacity, so to scale u_2 from zero to one, we set

$$u_2(50, 80, 100) = 0 \quad (11)$$

and

$$u_2(130, 200, 250) = 1. \quad (12)$$

The concepts of preferential and utility independence introduced previously were verified for the three capacity attributes X_2^{75} , X_2^{85} , and X_2^{95} . Thus we knew from the theorem stated previously that either

$$u_2(x_2^{75}, x_2^{85}, x_2^{95}) = \sum_{j=75,85,95} c_j u_2^j(x_2^j) \quad (13)$$

or

$$C u_2(x_2^{75}, x_2^{85}, x_2^{95}) + 1 = \prod_{j=75,85,95} [C c_j u_2^j(x_2^j) + 1], \quad (14)$$

where the u_2^j 's are the utility functions over X_2^j 's assessed on a zero to one scale as illustrated in Figure 4, and C 's and the c_j 's are scaling constants. Notice that the forms of (13) and (14) are analogous utility

functions expressed in (4) and (5). Since the following discussion concerns how the k_i 's and K 's in (5) are assessed, we will not indicate how the c_i 's and C 's in (14) were assessed, as the procedures are identical.

□ **Assessing the k_i scaling factors.** To illustrate the technique for assessing the k_i scaling factors, let us take cost (X_1) as an example. We asked the decision makers to compare a consequence with cost at its most preferred amount, and all the attributes at their least preferred amount, to a lottery yielding the consequence with all attributes at their most preferred amount with probability p or the consequence with all attributes at their least preferred amount with probability $1 - p$. The object is to find the value of p , call it p_1 , such that the decision maker is indifferent between the lottery and the consequence. Then, by using $u(\underline{x})$ from either (4) or (5) and equating expected utilities it is easy to show that k_1 must equal p_1 .

Using this procedure and then employing some consistency checks involving tradeoffs of two attributes at a time with all other attributes held constant, we ended up using the values of k_i indicated in Table 1.

□ **Assessing parameter K .** Since the sum of the k_i is 1.89, we know the utility function is multiplicative rather than additive; it is additive only if $\sum k_i = 1$. Therefore the value of K in (5) must be determined by evaluating it at $(x_1^*, x_2^*, \dots, x_6^*)$ where x_i^* the most preferred amount of X_i . This gives us

$$Ku(x_1^*, x_2^*, \dots, x_6^*) + 1 = \prod_{i=1}^6 [Kk_i u_i(x_i^*) + 1], \tag{15}$$

but from our scaling conventions, we know both $u(x_1^*, x_2^*, \dots, x_6^*) = 1$ and the $u_i(x_i^*)$'s are all one so

$$K + 1 = (Kk_1 + 1)(Kk_2 + 1) \dots (Kk_6 + 1). \tag{16}$$

Since the k_i are known, parameter K can be evaluated by solving (16). Since $\sum k_i > 1$, the value which K must assume is the solution

to (16) such that $-1 < K < 0$.¹² Using the k_i values from Table 1, we found $K = -0.69$.

□ **The utility function.** Procedures identical to those just illustrated were used to evaluate the c_i and C in (14). It was found that $c_1 = 0.3$, $c_2 = 0.5$, $c_3 = 0.4$, and $C = -0.46$. These parameters, together with those in Table 1 and $K = -0.69$, and the utility functions illustrated in Figures 3 and 4 represent the information necessary to specify the utility functions $u(x_1, x_2, \dots, x_6)$. The next section describes how it was used.

■ A computer was used to calculate expected utilities of the alternatives. To keep the calculations at a reasonable number, as mentioned earlier, many alternatives were eliminated before going through expected utility calculations. For instance, since military aircraft repre-

TABLE 1
SCALING FACTORS

ATTRIBUTE	k_i
$X_1 \equiv$ COST	0.48
$X_2 \equiv$ CAPACITY	0.6
$X_3 \equiv$ ACCESS TIME	0.10
$X_4 \equiv$ SAFETY	0.35
$X_5 \equiv$ DISPLACEMENT	0.18
$X_6 \equiv$ NOISE	0.18

5. The analysis

¹² As shown in [4].

sent a relatively insignificant amount of the total air traffic, alternatives differing only with respect to the airport for military operations were not considered separately. Secondly, alternatives which shifted certain types of aircraft from the Texcoco site to Zumpango and back again at a later date were excluded.

□ **Effective strategies.** Of the alternatives evaluated using expected utility, the top ten are indicated in Table 2. In the table, the expected

TABLE 2
THE TEN BEST ALTERNATIVES*

RANK	1975		1985		1995		EXPECTED UTILITY
	Z	T	Z	T	Z	T	
1	D	IMG	ID	MG	ID	MG	91.23
2	IDMG		IDMG		IDMG		90.90
3	I	DMG	ID	MG	ID	MG	90.79
4	ID	MG	ID	MG	ID	MG	89.30
5	ID	MG	IDMG		IDMG		88.10
6	ID	MG	ID	MG	IDMG		86.75
7	I	DMG	IDMG		IDMG		86.55
8	IG	DM	IDMG		IDMG		86.19
9	DG	IM	IDMG		IDMG		86.17
10	D	IMG	IDMG		IDMG		85.60

*TO HELP READ THE TABLE, THE ALTERNATIVE RANKED 1 IS DOMESTIC AIRCRAFT AT ZUMPANGO AND INTERNATIONAL, MILITARY, AND GENERAL AIRCRAFT AT TEXCOCO IN 1975, AND DOMESTIC AND INTERNATIONAL AT ZUMPANGO AND MILITARY AND GENERAL AT TEXCOCO IN 1985 AND 1995

utilities are calculated on a scale from zero to 100, where zero utility was assigned to a hypothetical alternative generated by taking the least desirable probability distribution for each attribute from the set of alternatives. Roughly speaking, this is the worst one could do. A utility of 100 was assigned to a hypothetical alternative generated from the most desirable probability distribution for each attribute. On this scale, the alternative of keeping all aircraft in Texcoco in all three years has an expected utility of 5.20.

It is clear that two types of strategies are effective. One might be categorized as the “all Zumpango” alternative and represents building a major new airport at Zumpango as soon as possible. The alternatives involving both international and domestic aircraft operating at Zumpango in all three years make up this category. The other type of effective strategy is the “phased development at Zumpango” characterized by either international or domestic aircraft operating in Zumpango in 1975 and then both by 1985 and 1995. Strategies which included keeping a part of the international or domestic traffic operating out of Texcoco through 1985 do not appear competitive in terms of effectiveness with the two types of strategies outlined above.

□ **The input-output display.** A graphic input-output display was developed to provide flexibility in using the model. Mainly, it was used as an efficient and accessible system for sensitivity analysis. It also permitted interested parties, other than the SOP to use the model if they wished. Input-output consoles were installed in various key

positions, including the President's office and used for communicating results of the analysis.

The input-output system allowed one to alter either or both of the probability or utility estimates for any specified alternative or alternatives. If it was desired to see the official SOP inputs before altering, this was done. Otherwise one could enter his own estimates without seeing any others.

To help in changing the utility function a number of questions had to be answered. Then a multiplicative or additive utility function was chosen to be consistent with the answers. It differed from the other utility function by having different scaling factors k_i . The probability density functions were altered by changing the maximum and minimum values for each of the attributes. The shapes of the marginal probability density functions or the single attribute utility functions could not be changed by graphic input-output. These changes required adjustments in the programs. Although they are important, they represent fine tunings relative to the alternatives provided for graphically.

□ **Use of the analysis.** The original purpose of the work described here was to identify effective strategies—as measured by our six measures of effectiveness—for developing the airport facilities of Mexico City. It was not to indicate what action should be taken by the Government of Mexico in 1971 to meet their needs. Only after the “effective strategies” had been identified, did we address this second question.

The alternatives in this study consisted of specifying various initial levels of development at Texcoco and Zumpango in both 1971 and 1975 at the end of the current administration's term in office. The alternative chosen in 1975 depends on critical events (such as technological developments, changes in demand, effects of air disasters, and changes in the public's preferences concerning ecology and priorities for government services) as well as on the original option chosen in 1971. Measures of effectiveness include “effectiveness” as defined in this paper, as well as political effects, flexibility to adapt the decision, and externalities such as regional development, etc.¹³

■ The conclusions fall into two categories: those pertaining to the study described in this paper and those relating to the analysis mentioned in the last section.

This study explicitly included the important factors relating to the “effectiveness” of various airport developments. These factors covered a rather broad spectrum of the overall impact, including effects on airport users, airport operators, and the government and public. The study also incorporated the uncertainty as to what consequences would eventually result from any specified course of action.

The quantification of preferences for the multiple attributes of this problem apparently represents the first attempt to obtain a complete utility function over several attributes and to use it in making a strategic policy decision. No unverified assumptions were needed to do this. Both the utility independence and preferential inde-

6. Conclusions

¹³ Details of this analysis are in de Neufville and Keeney [1].

pendence assumptions were empirically verified with various members of the SOP before assessing the utility function.

The probabilistic independence assumptions, on the other hand, were not empirically verified, and in fact, are probably an over-simplification as indicated in the paper. The assumptions were acceptable in this case for two reasons. First, the sensitivity analysis done by simultaneously altering the ranges of the marginal probability density functions indicated that the implications of the study were not very sensitive to the probabilities. Second, the use of this model was to *indicate* effective strategies. This information was included as one component part of the dynamic analysis. In the judgment of the SOP and the author, it is very unlikely that the implications of the dynamic analysis of alternatives to begin in 1971 would have changed if probabilistic dependencies among attributes were identified and assessed accordingly.

□ **Impact of the results.** The analysis described here is part of a broader study which played a role in the decision-making process of the Mexican Government, faced with alternatives for developing its airport facilities. From the 1965–1967 study of the SOP, which recommended moving the International Airport to Zumpango as soon as possible, and from the indications early in this study, it was clear that some members of the SOP still considered a major move to Zumpango the most effective strategy.¹⁴ This analysis, using the SOP's own estimates and preferences, indicated that a phased development involving a gradual shift toward Zumpango appeared equally good. This had a profound impact on many individuals within the SOP. Once other factors not included in this analysis (such as political considerations, flexibility of the policy, and externalities) were accounted for along with effectiveness in the dynamic analysis of alternatives open to the government in 1971, it was evident that the “phased development at Zumpango” policy was better than an “all Zumpango” policy.¹⁵

The SOP recommended a “phased development” strategy to the President in December, 1971. This represented a major change in their posture from the 1967 study. The 1971 recommendation accounted for critical events which might occur, and thus represented a dynamic strategy. This in itself was extremely important. The previous recommendations of the SOP were for master plans specifying what should be done at various points in time over the next 30 years. They did not plan present developments in light of the occurrence of possible future events, and essentially locked the government into a certain strategy.

References

1. DE NEUFVILLE, R. AND KEENEY, R. L. “Use of Decision Analysis in Airport Development for Mexico City,” Chapter 23 in *Analysis of Public Systems*, Drake, A. W., Keeney, R. L., and Morse, P. M., eds., Cambridge: M.I.T. Press, 1972.
2. HOWARD, R. A. “The Foundations of Decision Analysis.” *IEEE Transactions on Systems Science and Cybernetics*, Vol. SSC-7 (September 1968).

¹⁴ In [9] and [11].

¹⁵ In [1].

3. IPESA CONSULTORES AND SECRETARIA DE COMUNICACIONES Y TRANSPORTES. "Estudio de Ampliacion del Aeropuerto Internacional de la Ciudad de Mexico." Mexico, D. F., October 1970.
4. KEENEY, R. L. "Multiplicative Utility Functions." Technical Report 70, M.I.T., Operations Research Center, March 1972.
5. ———. "Utility Functions for Multiattributed Consequences." *Management Science*, Vol. 18, No. 5 (January 1972), pp. 276–287.
6. RAIFFA, H. *Decision Analysis: Introductory Lectures on Choices under Uncertainty* Reading, Mass.: Addison-Wesley, 1968.
7. ———. "Preferences for Multi-Attributed Alternatives." RM-5868-DOT/RC, The Rand Corporation, April 1969.
8. SCHLAIFER, R. O. *Analysis of Decision under Uncertainty*. New York: McGraw-Hill, 1969.
9. SECRETARIA DE OBRAS PUBLICAS. "Aeropuerto Internacional de la Ciudad de Mexico." Mexico, D. F., September 1967.
10. VON NEUMANN, JOHN AND MORGENSTERN, O. *Theory of Games and Economic Behavior* Second ed. Princeton: Princeton Univ. Press, 1947.
11. WILSEY Y HAM DE MEXICO, S. A. DE C. V. "Aeropuerto Internacional de la Ciudad de Mexico." Mexico, D. F., July 1967.

Pricing decisions of regulated firms: a behavioral approach

Paul L. Joskow

Assistant Professor of Economics
Massachusetts Institute of Technology

Once the prices for utility service are set by state regulatory authorities, they remain at fixed levels until they are officially increased or decreased by the regulatory agency. A utility company is not free to vary the prices of its service independently. This paper presents the results of an attempt to specify and estimate a behavioral model of the pricing decisions of regulated firms. The "threshold" behavior of regulated firms with regard to their decisions to seek general price increases from state regulatory authorities and their decisions to file "voluntary" price decreases is discussed, and two decision equations are specified and then estimated for regulated electric utilities in New York State. The decision to file for a price increase is shown to depend on the growth rate of earnings per share achieved by the firm in the current and previous year, on the level of interest coverage realized in the current year, and on a variable which measures prior expectations of success in the hearing room. The decision to file a voluntary price decrease is shown to depend on the growth rate in earnings per share and a variable which measures the expectations of "forced" regulation.

1. Introduction

■ Recent theoretical research has analyzed the behavior of a neo-classical profit-maximizing monopoly firm subject to a fair-rate-of-return constraint, where the fair rate of return lies somewhere between the profit maximizing rate of return and the market cost of capital.¹ Although the static allocational implications of this type model have been well described, only passing recognition has been given to the fact that in reality state regulatory agencies set the prices of service and not rates of return. Once the prices for utility service are set by the state regulatory authority, they remain at fixed levels until they are officially increased or decreased by action of that regulatory authority. A utility company is not free to vary the price of its service independently. As a result, an analysis of the interaction between regulated firms and their regulators and of the effects of the regulatory process on pricing decisions of regulated firms is of considerable interest.

Paul L. Joskow received the B.A. in economics from Cornell University (1968) and the M.Phil. and the Ph.D. in economics from Yale University (1970 and 1972, respectively). An assistant professor of economics at the Massachusetts Institute of Technology, he is currently conducting research in regulation in the insurance industry and in the economics of the electric power industry.

The author would like to thank Alvin Klevorick, John McGowan, and Richard Nelson for reading earlier drafts of this paper and for providing valuable criticism and encouragement.

¹ See Averch and Johnson [2].

This paper presents the results of an attempt to specify and estimate a model of the pricing behavior of regulated monopolies. The behavior of regulated firms with regard to their decisions to seek price increases from state regulatory authorities and their decisions to file “voluntary” price decreases is analyzed and two decision equations are estimated for regulated electric utilities in New York State.

The methodological underpinnings of this paper are in the tradition of the “behavioral theory of the firm” literature.² Although public utility firms probably try to do the “best they can” in some long-run sense, the complexity of the world that these firms are forced to operate in appears to lead them to use relatively simple short-run behavioral rules for making decisions about price changes. Research in the behavioral theory of the firm tradition argues that many firms in fact employ “rules of thumb” in making price and output decisions, based on indicators that are readily observable.³ To the extent that such rule of thumb decision making is a factor in firm behavior in general, it should be especially prevalent in monopoly industries protected from direct competition and the possibility of entry.

The author is aware of the varied criticisms that have been made of attempts to construct models of actual decision-making processes. Friedman⁴ argues that it is not a function of economic theory to recreate the real world, but to construct theoretical paradigms that predict well. Current models of regulated firms do not predict pricing behavior very well. In particular they do not predict the lack of responsiveness of prices to rising taxes during the early and middle 1960s nor to rising costs during 1967 and 1968, nor do they explain the plethora of formal rate cases in the United States during the period 1961–1968 when public utility profits were rising.⁵ In general, these models have restricted the kinds of questions asked to efficiency questions (primarily static production efficiency) and are not very useful for describing price and output dynamics of regulated firms.⁶ In addition, the value of “as if” models declines as we not only be-

² Cyert and March [4], p. 12, and Simon [15].
³ Cyert and March [4], p. 12.
⁴ In [5].
⁵ Electric and Gas Utility Rate Cases Decided by State Commissions 1961–1970:

<u>year</u>	<u>number of cases</u>
1961	23
1962	18
1963	8
1964	6
1965	6
1966	6
1967	6
1968	12
1969	32
1970	41.

Notice the tremendous increase in numbers of cases as inflation forced firms to come to the Commission for rate increases. (Source: [17].)

⁶ For a fascinating discussion of how different model paradigms affect what we see, the questions we ask, and the public policy implications of a particular decision-making process, from the viewpoint of a political scientist, see Allison [1].

come interested in predicting how firms behave given current structural interrelationships, but begin to ask questions about structural changes aimed at changing the nature of firm responses. For those interested in public policy analysis regarding regulated industries, a more detailed understanding of firm decision processes, decision processes of regulatory agencies, and their interrelationship appears to be in order.

A more telling criticism of the behavioral theory of the firm methodology is that results obtained are not general nor can analytical behavioral models be constructed. This study attempts to respond to this type of criticism in a number of ways. A general probability model of firm pricing decisions is specified. This model appears to describe the short-run pricing decision process of regulated firms fairly well, at least on an appreciative level. Particular decision variables are then specified and inserted into the general probability model, and finally two decision equations are estimated, based on observations on the pricing behavior of a set of regulated electric utilities. Both the general model specification and the particular variables chosen are largely derived from detailed case study work which the author has undertaken, involving a small number of decision makers, to gain insights into the decision-making processes of regulated firms, their regulators, and the interaction of the two over time.⁷ The estimation, however, is based on observations on the price behavior of several firms, most of which were not themselves the subject of case study work. The approach taken here sacrifices some of the specificity of the decision rules derived by Cyert and March for a department store⁸ in an effort to obtain a more general model which predicts well for a large number of similar firms and also provides the opportunity for general analytic work.

2. Pricing decisions of regulated firms

■ Any corporate entity must make a large number of decisions each year regarding such things as output prices, quantity produced, technology, etc. A regulated public utility is unique with regard to its ability to adjust prices, especially to make price increases. Price adjustments cannot be taken unilaterally by the firm, but require the approval of the regulatory authority. For a regulated firm, a decision must be made periodically regarding the general level of the company's rates and whether or not the company should go to the commission to get its rates changed. Consider first the decision to apply for a rate increase.

As costs increase with prices fixed, the declining profitability of the company's operations will at some point prompt it to go to the regulatory commission to ask for a rate increase. Such a request will generally prompt a rather lengthy public hearing during which recent company operating experience will be reviewed, testimony from interested intervenors will be taken, and a decision concerning how large a rate increase will be allowed (if any) decided upon.⁹ Firms hesitate to go to the commission for a rate increase every time some component of costs increases for a number of reasons: many firms

⁷ See Joskow [7, 8].

⁸ Cyert and March [4], pp. 128–148.

⁹ See Joskow [8].

appear to be able to earn substantially more than their cost of capital without being forced to cut prices¹⁰ (this was especially true during the period 1960–1967). The promulgation of a rate hearing aimed at adjusting prices upward for, let us say, a tax increase could result in a *price reduction* to bring earned rates of return into conformity with prevailing allowed rates of return. In addition, as a result of regulatory lag and other informal regulatory constraints, once a rate increase has been applied for, the possibility of another rate increase is effectively eliminated for 18 to 24 months. Finally, formal regulatory hearings subject the operations of the firm, in terms of financial position, output strategies, and service quality, to a degree of scrutiny which it would not normally have to undergo. The confrontation with intervenors representing the interests of various members of the public appears to be especially troublesome. The decision to apply for an increase or not appears to be a big decision, taken separately from the secondary decision of determining exactly how much to ask for if an increase is in fact requested. The question decision makers must answer has a dichotomous solution: file an increase request or not. In response to this rather complex situation, regulated firms appear to look to a fairly small set of indicators in determining whether the status quo prices or a requested rate increase will prevail.

The status quo rate position appears to be a position of price stickiness. Under certain circumstances firms will tolerate fairly substantial movements in overall corporate earnings as well as movements in earnings of particular departments (gas or electric for combination companies) without in fact being pushed to the decision to file for a price increase. For example, many utility companies absorbed rather substantial increases in property taxes during the early 1960's without making any attempt to pass them on through higher prices. Individual firms appear to have target values for a number of financial variables. The decision to file or not to file for a rate increase appears to be a function of deviations of actual performance from target performance and an assessment of the chances of getting favorable action from the regulatory authority.

Specifically we may view each firm as having a number of financial

¹⁰ Litzenberger and Rao [12].
Although many firms were observed earning more than their cost of capital, my belief is that this result emerges because most of these firms did not have to come to the FPC for rate increases during the period. In its starkest form, my hypothesis is that regulatory commissions only regulate the profits of firms attempting to raise prices. If we compare earned rates of return of firms which did not seek to raise their prices with those that did, we should find that the former group is able to earn much higher profits than the latter. I am in the process of collecting data for such a study; the following data are available now for New York State.

Average Earned Rate of Return of Companies Seeking Rate Increase in Current or Previous 2 Years		Average Earned Rate of Return of Companies Not Seeking Rate In- crease in Current or Previous 2 Years
<i>year</i>		
1962	5.63%	6.65%
1963	5.50	6.64
1964	5.18	6.73
1965	5.35	7.05

(Source: Federal Power Commission, *Statistics of Class A&B Electric Utilities*, published annually.)

indicators which it looks at each period. Let X_i be the value of the i th financial indicator employed as a decision variable in a particular period. Firms will also have a target value for each of the financial indicators. Let $\bar{X}_i$ be the target value of the i th financial indicator. In addition, firms will have a perception of the chances of getting a rate increase approved by the Commission. Let P_s be the perception of the chances of getting a rate increase approved by the regulatory Commission if a request is made. Finally we specify a linear decision index (I_k) which is a linear function of these variables:¹¹

$$\begin{aligned} I_k &= \sum \alpha_i (\bar{X}_i - X_i) + \beta P_s \\ &= \sum \alpha_i \bar{X}_i - \sum \alpha_i X_i + \beta P_s \\ &= \gamma - \sum \alpha_i X_i + \beta P_s \\ &= \gamma + \sum (-\alpha_i) X_i + \beta P_s. \end{aligned}$$

In any year if the value of I_k is greater than some "threshold value" $\bar{I}_k$ the firm will file for a rate increase; otherwise it will not:¹²

$$I_k > \bar{I}_k \text{ go for rate increase (INC} = 1)$$

$$I_k \leq \bar{I}_k \text{ do not go for rate increase (INC} = 0),$$

where the *ex post* behavior of a particular firm is represented by the variable *INC* which takes on the values zero or one.

Finally the "threshold values" of the decision index that will trigger a rate increase appear to be different for different firms. We assume here that the threshold values are distributed normally over the population of firms with mean zero and standard deviation unity.¹³ For a randomly chosen firm we can thus specify the *ex ante* probability that a rate increase will be filed for given levels of the independent variables as the following:

$$P(\text{INC} = 1) = P(I_k > \bar{I}_k | X_1 \dots X_n, P_s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\bar{I}_k} e^{-u^2/2} du \quad (1)$$

= Probability that a rate increase will be filed by a randomly chosen firm for given levels of the independent decision variables.

Equation (1) turns out to be in the form of the multivariate probit analysis model. This model will be further discussed below before we estimate the parameters of the decision function using observations on the past behavior of a set of regulated firms.

Casual observation of the recent behavior of regulated public utilities appears to indicate that all public utility firms do is request

¹¹ Choice of a linear decision index as opposed to some nonlinear form is completely arbitrary. Case studies only revealed which variables decision makers looked at and notions of which variables received more weight but not a particular functional form determining exactly how the weighting of the variables was done.

¹² This is the "satisficing" threshold of the firm.

¹³ The choice of a normal distribution is arbitrary. However, a symmetric distribution seems to be an appropriate one to look at first. The analysis would be exactly the same with any probability distribution specified, although other distributions would make parameter estimation much more difficult. The choice of mean zero and standard deviation unity are not crucial, but serve only as scale factors. Predicted probabilities would be exactly the same if some other mean and standard deviation were chosen.

price increases as often as they possibly can. If we look back to the periods of the late 1950s and early and middle 1960s, however, we observe some price increases, many instances in which firms do not change their prices in a particular year, and some situations in which price reductions are observed. Price reductions in general appear to be "voluntary" and occur without a formal regulatory hearing's taking place. Although the firm's decision to decrease rates appears to be mathematically analogous to the decision to increase rates, a discussion of the motivations for price decreases is necessary.

Once the price level for the firm has been set by a regulatory authority, the firm experiencing growing demand and declining average costs will find its profits growing from period to period. I previously argued that to the extent that a strict regulatory constraint exists, it exists only for those firms actually engaging in the formal hearing process by seeking rate increases. Firms which can increase their profits without increasing prices will tend, within rather wide boundaries, to be left alone by the regulatory authority (they will not in general be called in for a forced regulatory review and a forced price reduction). As profits increase for these firms, there may be an increasing tendency for the staff of the regulatory authority to "suggest" price reductions or improvements in service, often in response to external pressures from citizens' groups or politicians. Thus, although there is a constraint tending to limit continual growth of profits, it appears to be a much looser kind of constraint than much of the theoretical literature might lead one to believe. Regulated firms do appear to respond to this moral suasion from time to time, because the regulatory authority has the power to force a formal regulatory review¹⁴ and firms feel that they will do better in the long run if they keep the commission happy by filing a rate reduction from time to time rather than by waiting for the commission to come after them.¹⁵

Mathematically the firm's decision to decrease rates is similar to its decision to increase rates. It continues to possess targets for certain financial variables (although the variables it looks at in making a decision about price reductions may be different from those it uses for making price increases), and also has a perception of the chances that the commission will initiate a regulatory hearing and a forced reduction in rates. We may call this "forced regulation" variable P_D . The firm's decrease decision index can then be represented by

$$D_k = \delta + \sum g_i X_i + \xi P_D$$

$$D_k > \bar{D}_k \text{ file a voluntary price decrease } (DEC = 1)$$

$$D_k \leq \bar{D}_k \text{ do not file a voluntary price decrease } (DEC = 0),$$

where the threshold values of the decrease decision are assumed to be normally distributed with mean zero and standard deviation unity over the population of firms and where DEC characterizes *ex post* dichotomous decrease behavior taking on the values zero or unity.

The probability that a randomly chosen firm will file a voluntary

¹⁴ See *Re Rochester Gas And Electric Company*, 43 PUR 3d, where a forced regulatory review was initiated.

¹⁵ See note 10.

price decrease given the values of the exogenous variables is thus

$$P(DEC = 1) = P(\bar{D}_k > D_k | X_1 \cdots X_n, P_D) \\ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{D_k} e^{-u^2/2} du. \quad (2)$$

Equations (1) and (2) make up the two pricing decision functions employed by regulated firms. Notice that the increase decision and the decrease decision are treated as autonomous decisions in this model and the “do nothing” possibility is then treated as a residual:

$$P(NOTH = 1) = 1 - P(DEC = 1) - P(INC = 1). \quad (3)$$

This means that all behavior not explained by either of the first two equations falls automatically into equation (3). Problems which may be engendered by this treatment are discussed after the decision variables are specified and the maximum likelihood estimating technique employed is described.

3. Specification of the decision variables for estimation

■ The problem which must be dealt with now is the specification of the actual variables regulated firms in fact look at in making their decisions, and estimation of the parameters of each of the decision equations. Variable specification will be dealt with first.¹⁶

□ **The decision to file for a rate increase.** The firms that are being discussed here are combination companies: one corporate entity operating two or more service departments—usually electric, gas, and steam. Rate increases for each of the operations are normally made and considered separately although they are interconnected to the extent that the performance of each department affects the overall performance of the firm.

The financial officers of the firm appear simultaneously to survey the overall financial position of the firm and to search over the various operations of the firm to see where a rate increase might most easily be obtained. They must assess the return each department (gas, electric, or steam) is earning relative to the most recent activities and decisions of the commission to see whether or not they have a good chance of getting a rate increase approved in that department.

The financial variables chosen were those which utility decision makers mentioned as being most important in their decision to change prices from their existing level. The financial indicator that seems to be of most interest to utility people is the growth rate in earnings per share, and this indicator is used in both the increase and the decrease decision equations.

Since firms seem to be adverse to entering formal hearings unless they must absolutely do so, it seems only natural that they should take a long-term view of their financial picture and not respond too quickly to a “bad” period that might be made up in the future at existing rates. Although the case study indicated this inclination, exactly how many periods of performance firms use to establish a view of their

¹⁶ This specification is based on a case study conducted by the author which entailed discussions with some utility decision makers and public utility commissioners. See Joskow [7], Chapter II, pp. 71–80.

financial picture was not revealed. It appears to be more than a year, however.

Since the model presented here is based on annual observations, and since firm decision makers get detailed financial data at least every quarter, a two-year time horizon (eight quarters) seemed to be a reasonable amount of time for an assessment of whether the firm was actually "stuck" in a position of poor financial performance at existing rates. To take this behavior into consideration, the growth rate in earnings per share and the growth rate in earnings per share lagged one period are used as explanatory variables in both the increase and decrease equations.

The second important variable examined by utility decision makers in making the decisions to increase rates or not is the level of interest coverage the firm is achieving.¹⁷ Interest coverage considerations become especially important when the coverage level becomes so low that bond ratings are affected or the minimum interest coverage level (mandated by the firm's first mortgage indentures) is approached. The level of interest coverage experienced by the company as a whole is therefore included as a decision variable in the price increase decision index.

A variable to serve as an indicator of the chances of success of the prospective rate increase is also required. Firms seem to get information about the behavior of the regulatory commission by observing its recent activities with regard to other firms under the commission's jurisdiction. In particular, a variable which compares the rate of return the department is earning with the rate of return allowed by the commission in the most recent case before the commission, seems to be appropriate.

By looking at a recent decision the firm obtains information about the commission's current thinking concerning what the appropriate rate of return for a utility department is at prevailing rates of interest. The firm can compare the rate of return each of its departments is earning with the returns recently allowed in an attempt to ascertain whether the commission would view the return being earned by one of its departments as being low enough to justify a rate increase. If a particular department is earning a rate of return substantially below that allowed in a recent case, the firm should view this as an indication that the rate increase for this department should have a very good chance of getting approved.

The variable chosen is a simple quotient of the rate of return the department under consideration actually earned in the most recent financial period and the rate of return the commission allowed in the most recent formal rate case:

$$P_s = ROR_e / ROR_1,$$

where

ROR_e = rate of return earned by (electric or gas) department;

ROR_1 = rate of return allowed in most recent formal rate case.

In its simplest form then, the decision to increase rates or not for a particular department depends on the growth rate in earnings per

¹⁷ Interest Coverage = Gross Income Before Income Taxes Divided by Total Interest Charges.

share and the growth rate in earnings per share lagged one period for the company as a whole, the interest coverage for the company as a whole, and a variable equal to the quotient of the rate of return earned by the department under consideration and the rate of return allowed by the commission in the most recent rate case:

$$P(INC = 1) = P(I_k > \bar{I}_k) \quad \bar{I}_k \sim N(0, 1)$$

and

$$I_k = \gamma + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \beta P_s, \quad (1a)$$

X_1 = growth rate in earnings per share,

X_2 = growth rate in earnings per share lagged one period,

X_3 = level of interest coverage,

$P_s = ROR_s/ROR_1$, and

$INC = 1$ if an increase is filed by a particular firm
 $= 0$ otherwise.

The institutional factor of regulatory lag must be taken into consideration. Since there may be a lag of more than a year between the time a firm decides to request a rate increase and the time a final decision is made, we would expect that if financial indicators during the current period indicated the necessity for a rate increase, these financial indicators will continue to indicate the need for a rate increase during the period that the request is pending. We do not observe, however, a rate increase request during this lag year, since one is already pending.

Regulatory lag has been incorporated in two different ways. In the first formulation of the model (1a), observations for the lag years are omitted because the probability of a petition for a rate increase is zero regardless of what the financial variables indicate. This follows from the institutional constraint that when a case is pending we cannot observe another request. Firms observed are characterized as having filed for a rate increase ($INC = 1$) or having done otherwise ($INC = 0$) and lag years are omitted.

The second formulation (1b), treats the independent variable as the decision to increase rates *or* to leave a previous petition for a rate increase pending. Model (1b) assumes that the lag years will continue to result in financial performance that would indicate a rate increase request if one were not already pending. Model (1a) predicts the probability of a rate increase conditional on the fact that a rate increase request is not already pending, while model (1b) predicts the probability that a rate increase request will be made *or* a request is pending. The model formulation for (1b) is exactly the same as for (1a) except firms are characterized as filing for a rate increase or having one pending ($INC^* = 1$) and having done otherwise ($INC^* = 0$).

□ **The decision to make “voluntary” rate reductions.** The objectives of the firm in deciding whether to reduce the prices it charges for utility service are, of course, very similar to those associated with the decision to petition for rate increases. The financial officer is interested in maintaining at least some target level of financial performance. At the same time he must make some kind of assessment of

whether the earnings of any of the departments of the utility are so high as to prompt the commission to “suggest” an appropriate rate decrease or even to prompt it to institute a formal proceeding.

The growth rate in earnings per share and the growth rate in earnings per share lagged one period are used as financial indicators. Once again, it is expected that firms look at more than one period of financial performance in making their decision to decrease rates. Two periods of financial data give firms more information concerning the soundness of their financial position at current rate levels. They would not want to decrease rates if good financial performance were only transitory since they would have to go back to the commission for a rate increase once performance returned to normal. Interest coverage considerations do not appear to be a consideration on the decrease side, since when a decrease is being considered, interest coverage levels are far in excess of their danger points. This variable has therefore not been included in the decrease decision equation.

An indicator of the chance that the commission will either suggest or impose a rate decrease must be constructed. Once again firms appear to form their perceptions about how much is “too much” by observing interactions between the commission and other firms. Since firms would like to earn as much as they can, but also avoid “forced” profit reductions by the regulatory commission, they must have some indicator of how much profit is too much profit in the eyes of the commission. The variable constructed as a measure of the chances of drawing the wrath of the commission, or an indicator of how much is “too much” is the following: the quotient of the rate of return that the department (electric or gas) is actually earning and the highest rate of return the commission let stand for a comparable electric or gas department or company in the state in the previous period:

$$P_D = ROR_e / ROR_h,$$

where

ROR_h = the highest rate of return the commission let stand for a comparable (electric or gas) utility in the previous period.

The interaction between the firm and the commission is accounted for by this variable, while the financial performance target of the firm itself is accounted for by the first two variables. By using only a one-year horizon on the rate of return earned by the relevant department of the firm, it is being assumed that when rates of return do get too high the commission will bring quick action to have prices reduced.¹⁸

In a world with stable capital costs (interest rates) we would be inclined to look not just at the previous periods’ peak earnings, but at some average over past periods’ peak earnings in an effort to judge how much is too much. In a world where capital costs fluctuate, however, the most recent figure gives the firm an idea of how much is “too much” in light of prevailing interest rates. The model for the firm’s decision to decrease rates for a particular department or to do

¹⁸ The department’s current rate of return is compared with the highest rate allowed in the previous period because when the company is making its decision, data for the other companies for the most recent financial period are not yet available to it.

otherwise is therefore:

$$P(DEC = 1) = P(D_k > \bar{D}_k) \quad \bar{D}_k \sim N(0, 1) \quad (2)$$

$$D_k = \delta + g_1 X_1 + g_2 X_2 + \xi P_D,$$

where X_1 and X_2 are defined as before, $P_D = ROR_e/ROR_h$ and $DEC = 1$ when a voluntary decrease is made and $DEC = 0$ otherwise.

4. Estimation of the decision equations

■ Decision models (1) and (2) are in the form of the multivariate probit analysis model often used in biometric analysis and which has seen some use in economics.¹⁹ Nonlinear estimating packages are available for calculating the maximum likelihood estimates of the coefficients of each of the decision indexes (1a), (1b), and (2) and one such program has been used to obtain the estimates presented here. All that we need therefore is observations on the pricing decisions of a set of regulated firms, observations on the independent decision variables of these firms, data on allowed rates of return by the regulatory authority with jurisdiction over these firms and data on earned rates of return achieved by the departments under consideration. The empirical analysis concentrates on the pricing decisions of the electric departments of a set of combination companies in New York State and under the jurisdiction of the New York State Public Service Commission.²⁰

5. The data

■ The model presented above was estimated using data collected on the behavior of the seven major electric utilities in New York State. There are 63 observations available on seven combination companies for the period 1961–1969.²¹ The model has been estimated only for changes in electric rates because it is only for electric companies (or electric departments of combination companies) that consistent and meaningful rate-of-return data are available.²²

¹⁹ See especially Goldberger [6], pp. 250–251, and Tobin [16].

The probit model takes the following form:

$$Pr(E|C_1 \dots C_N) = S_N(\sum_i g_i C_i),$$

where $Pr(E|C_1 \dots C_N)$ is the conditional probability of the event's taking place, given the values of the variables $C_1 \dots C_N$ that affect the outcome of the event. Here S_N is the cumulative standard normal distribution function, $\sum_i g_i C_i$ is the weighted sum of the independent characteristic variables. The g_i 's are the parameters estimated from data using maximum likelihood techniques. The probit model restricts the calculated probabilities to the zero to one range.

²⁰ It is important to isolate companies which are subject to the same regulatory authority. Different commissions have different modes of behavior and calculate allowed rates of return differently. Although the structure of the model should be the same across states, allowed-rate-of-return figures are not in general comparable and particular parameters in the decision index may differ.

²¹ Problems arising due to the use of a time series of cross sections have largely been ignored. It is not clear what types of biases are introduced into a "probit" model when "firm effects" or "time effects" are not properly taken into account. All of the equations were estimated using firm dummy variables in an effort to unmask possible firm effects. None of the coefficients of these variables was significantly different from zero, indicating that firm effects may not be an important problem.

²² The earnings per share data come from *Moody's Public Utility Manual* and the rate-of-return data come from the FPC publication, *Statistics of Class A and B*

■ There is no way of telling exactly when a firm made its decision to petition for an increase or decrease in rates or exactly what financial information was available to it when it made its decision. The following timing conventions were adopted in estimating the model.

Rate increase requests occurring before June 30 of a particular year were associated with the characteristics of the previous year. This timing convention is used under the assumption that processing of data and the decision-making process itself take time so that information from the previous year was the most recent available when the decision was made. Similarly, rate increase requests occurring after June 30 were associated with that year under the assumption that a substantial amount of financial information about the current year was already available. For example if an increase request took place on June 29, 1963, the right side of equation (1a) would have EPS_{62} , EPS_{61} , COV_{62} , ROR_{62}/ROR_1 .

Rate decreases reported in the Annual Report of the New York State Public Service Commission were associated with the financial indicators of the previous year. This convention was adopted under the assumption that the PSC will bring pressure to decrease rates based on the results of an annual financial review.

■ The following signs and magnitudes for the maximum likelihood estimates are expected.

□ **Increase decision.** The increase equation is

$$\hat{I}_k = \hat{\gamma} + \hat{\alpha}_1 X_1 + \hat{\alpha}_2 X_2 + \hat{\alpha}_3 X_3 + \hat{\beta} P_s, \quad (1a)$$

where $\hat{I}_k$ = the calculated probit index and $(\hat{\cdot})$ indicates maximum likelihood estimates.

On the basis of the previous discussion we would expect the following signs for the coefficients of the probit model.

$\hat{\alpha}_1$: NEGATIVE. As the growth rate in earnings per share increases the probability of a rate increase request should become smaller.

$\hat{\alpha}_2$: NEGATIVE.

$\hat{\alpha}_3$: NEGATIVE. As the level of interest coverage increases, the firm has less need and less justification to file for a rate increase. The probability of making an increase request should therefore decrease as the level of interest coverage rises. We expect this result since as the level of interest coverage rises the firm should be moving further and further from its interest coverage constraint and therefore feel less pressure to raise rates to put it in a more favorable coverage position.

$\hat{\beta}$: NEGATIVE. Higher values of P_s indicate less likelihood of obtaining favorable action on a rate increase request. The probability of making a request for an increase will therefore decrease.

$$\hat{I}_k^* = \hat{\gamma}^* + \hat{\alpha}_1^* X_1 + \hat{\alpha}_2^* X_2 + \hat{\alpha}_3^* X_3 + \hat{\beta}^* X_4. \quad (1b)$$

(Case pending treated as $INC = 1$.)

Utilities (published annually). Information concerning general price increases and decreases comes from the Annual Report of the *New York State Public Service Commission* (1960–1970) and Public Service Commission dockets.

6. Timing problems

7. A priori expectations

Coefficients $\hat{\alpha}_1^*$, $\hat{\alpha}_2^*$, $\hat{\alpha}_3^*$, and $\hat{\beta}^*$ are expected to have the same signs as $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\alpha}_3$, and $\hat{\beta}$, respectively.

□ **Decrease decision.** The decrease equation is

$$\hat{D}_k = \hat{\delta} + \hat{g}_1 X_1 + \hat{g}_2 X_2 + \hat{\xi} P_D. \quad (2)$$

We would expect the following signs for the coefficients.

- $\hat{g}_1$: POSITIVE. As the growth rate in earnings per share increases, the probability of a rate decrease should also become larger.
- $\hat{g}_2$: POSITIVE. The firm has more than a one-year time horizon. Its financial performance in the recent past should be a significant factor in the firm's decision to decrease rates this year.
- $\hat{\xi}$: POSITIVE. As the rate of return being earned by the electric company approaches the highest rate of return being earned by an electric company in the state during the *previous* period, the probability of a rate decrease should increase.

8. Empirical results

■ The following tabulations present the maximum likelihood estimates of the probit model based on data for eleven electric utilities in New York, from 1961–1969.

Equation (1a) increase:

<u>coefficient</u>	<u>estimate</u>	<u>asymptotic standard error</u>	<u>asymptotic normal test</u>
$\hat{\gamma}$	14.47	5.47	2.64
$\hat{\alpha}_1$	— 0.14	0.06	—2.33
$\hat{\alpha}_2$	— 0.22	0.10	—2.20
$\hat{\alpha}_3$	— 0.71	0.47	—1.51
$\hat{\beta}$	—10.62	4.11	—2.58

Test that $\hat{\alpha}_1 = \hat{\alpha}_2 = \hat{\alpha}_3 = \hat{\beta} = 0$,

$$(-2.0) \cdot \log \text{ of likelihood ratio} = 30.78.$$

Degrees of freedom = 4.

Equation (1b) increase:

<u>coefficient</u>	<u>estimate</u>	<u>asymptotic standard error</u>	<u>asymptotic normal test</u>
$\hat{\gamma}^*$	10.10	2.85	3.54
$\hat{\alpha}_1^*$	— 0.12	0.05	—2.42
$\hat{\alpha}_2^*$	— 0.14	0.05	—2.70
$\hat{\alpha}_3^*$	— 0.66	3.18	—2.08
$\hat{\beta}^*$	— 6.62	2.31	—2.87

Test that $\hat{\alpha}_1^* = \hat{\alpha}_2^* = \hat{\alpha}_3^* = \hat{\beta}^* = 0$,

$$(-2.0) \cdot \log \text{ of likelihood ratio} = 43.35.$$

Degrees of freedom = 4.

Equation (2) decrease²³:

<i>coefficient</i>	<i>estimate</i>	<i>asymptotic standard error</i>	<i>asymptotic normal test</i>
$\hat{\delta}$	-7.80	2.85	-2.74
$\hat{g}_1$	0.13	0.05	2.69
$\hat{g}_2$	0.09	0.04	2.07
$\hat{\xi}$	6.29	2.89	2.18

Test that $\hat{g}_1 = \hat{g}_2 = \hat{\xi} = 0$,
 (-2.0)·log of likelihood ratio = 24.75.

Degrees of freedom = 3.

A likelihood ratio test was used to test the null hypothesis that all of the coefficients (except the constant) are zero against the alternative that the coefficients are different from zero in each equation. For example where L^* is the maximized value of the likelihood function:

$$\lambda = L^*(\gamma, 0 \dots 0)/L^*(\hat{\gamma}, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\beta}),$$

$$\log \lambda = \log L^*(\gamma, 0, 0, 0, 0) - \log L^*(\hat{\gamma}, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\beta});$$

and where $-2 \log \lambda$ is distributed as χ^2 with four degrees of freedom for equation (1a) for large numbers of observations.²⁴ The value of γ in this test is such that $S_N(\gamma) = \frac{I}{N}$ where I = number of increases (decreases) and N = number of observations.

The likelihood ratio test leads us to reject the null hypothesis that all coefficients are zero for large values of $-2 \log \lambda$ and the null hypothesis is rejected for all three decision equations at the 5-percent level.

The likelihood ratio test can also be used to test hypotheses on each of the coefficients (h_i) but this can involve a computational burden. Instead, tests of the individual coefficients were based on the fact that for large samples the maximum likelihood estimates of these coefficients are approximately normally distributed with means equal to the maximum likelihood estimates and variance-covariance matrix estimated by the inverse of the negative of the matrix of second partial derivatives of the likelihood function evaluated at the maximum likelihood estimates. Using this fact, the null hypothesis $h_i = 0$ can be tested quite simply against the alternative $h_i \neq 0$. Under the null hypothesis, for large samples, $\hat{g}_i/\hat{\sigma}_i$ is distributed approximately as a standard normal random variable. The null hypothesis is rejected for $|N_k| = \hat{h}_i/\hat{\sigma}_{h_i}$ large. This is the source of the significance tests in the three tabulations presented.

The results appear to give encouragement to the behavioral model postulated. All of the coefficients of equation (1a) have the the correct *a priori* signs. Coefficients γ , α_1 , α_2 , and β are significant at the 5-percent level. The coefficient of the coverage variable is not, however, quite significant at the 10-percent level.

²³ This equation was also estimated with a coverage variable included. The estimated coefficient of the coverage variable was insignificant with a variance larger than the estimated value of the coefficient.

²⁴ See Mood [13], p. 259.

As the growth rate in earnings per share declines, the probability of a rate increase request increases. Previous year earnings per share growth seems to have a larger impact on the decision, however, than does the current growth rate in earnings per share. As the earned rate of return declines relative to that allowed in a recent case, the probability of a rate increase also increases. As the level of interest coverage declines the probability of a rate increase request increases. This is the behavior that was predicted.

The results for equation (1b) are also quite good. All of the coefficients are significant at the 5-percent level and they all have the predicted signs. In addition, the significance levels of the coefficients for model (1b) are higher than for the corresponding coefficients of model (1a). This indicates that the independent variables may be better able to predict the probability of a rate increase request or case pending than the probability of a rate increase request conditional on a case not pending. The specification to be preferred will depend on the use to which the model is to be put.

The results for the decrease equation (2) are also encouraging. All of the coefficients are of the predicted sign and are significant at the 5-percent level. As the current or lagged growth rate in earnings per share increases, the probability of a rate decrease increases, while as the "risk" of being reviewed increases so does the probability of a rate decrease. Appendix 1 presents the calculated probabilities based on the probit models and compares the estimates with actual behavior, and Appendix 2 discusses regulatory lag further.

We note finally that the residual probability—the decision to do nothing—has in all cases a probability that lies in the zero to one range. For this sample at least, treating the decisions as being separate does not result in meaningless probabilities for the do-nothing decision.²⁵

²⁵ We can also examine the residual relationship, but it is a difficult relationship to evaluate:

$$(i) \widehat{P(NO\text{TH})} = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\hat{\Gamma}_k} e^{-u^2/2} du - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\hat{\Delta}_k} e^{-u^2/2} du \text{ Residual Relationship}$$

and

$$(ii) \frac{\partial \widehat{P(NO\text{TH})}}{\partial X_i} = \frac{-\hat{g}_i^{INC}}{\sqrt{2\pi}} e^{-\hat{\Gamma}_k^2/2} - \frac{\hat{g}_i^{DEC}}{\sqrt{2\pi}} e^{-\hat{\Delta}_k^2/2},$$

where X_i = i th independent variable,

$\hat{g}_i^{DEC}$ = estimated coefficient of probit decrease index, and

$\hat{g}_i^{INC}$ = estimated coefficient of probit increase index.

For variables that appear in one equation but not the other, changes in these variables result in a shift in probability between the decision equation and the residual equation ($\hat{g}_i^{DEC} = 0$, for example). The magnitude of the shift will depend on the value of the calculated probit index in the decision equation in which the variables appear and the size of the coefficient. The direction of the shift in probability will depend on the sign of the coefficient being varied.

If the variable being changed appears in both decision equations, the variation in this independent variable will in general affect the probability in both decision equations and in the residual equation. Since the variables shared by both decision equations have opposite signs and approximately the same absolute values, the level of the change in the probability of doing nothing will depend on the relative magnitudes of $|\hat{\Gamma}_k|$ and $|\hat{\Delta}_k|$. If these values are equal, changing an independent

■ The work that has been presented here has sought to make use of the behavioral theory of the firm methodology to gain insights into the decision-making process of regulated public utilities. The results appear to have both positive and methodological implications that are worth mentioning.

The first thing to note is the nature of the response of prices to changes in those things which affect the firm's financial position such as taxes and inflation. Whether or not a particular firm will respond to an increase in taxes or an increase in production costs by attempting to raise prices will depend on how close financial variables are to their targets and the "threshold value" of the decision index for the firm. During periods when financial performance is very good most firms will simply absorb the tax or cost increase and not attempt to pass it on to consumers through higher prices. Similarly in periods of very poor financial performance even small tax or cost increases may prompt firms to seek price increases in an effort to pass them on to consumers. The model therefore provides an explanation for why many public utilities absorbed continually rising property taxes during the early and middle 1960s, by and large without an attempt to obtain rate increases. In addition, it explains why there were so few formal regulatory cases during the early and middle 1960s, why there has been such a dramatic increase in the number of formal rate cases since 1968, and why utility firms are trying to get tax increases automatically rolled into their rates.²⁶

Much of the recent literature on public utility regulation has been concerned with the effects of rate-of-return regulation on production efficiency. Although it appears to make little sense to examine the technical efficiency implications of regulation without also examining the exchange efficiency implications²⁷ if *social* welfare judgments are to be made, this model also provides some insights into the incentives of firms to minimize production costs.

Essentially, the model implies two constraint regions. One is a lower boundary (a rate-of-return boundary). As firms approach this, the chances of getting a rate increase approved increase.²⁸ The second is an upper rate-of-return boundary region. As firms approach this, attempts are made to keep the earned rate of return from rising any further. Firms operating at rates of return between these boundaries²⁹ can effectively maximize profits without any constraint and have an incentive to minimize costs. It would appear that firms operating on either the lower or upper boundary do face some type of regulatory

variable will change the probabilities of an increase request and a decrease request by approximately the same amount (but in opposite directions) and leave the residual probability unaffected. As $|\hat{I}_k|$ and $|\hat{D}_k|$ become more unequal, the residual equation receives more of the impact of changes in shared independent variables.

²⁶ See *Re Consolidated Edison Company et al.* 77 PUR 3d.

²⁷ See Kahn [9] and Klevorick [10].

²⁸ This appears to be analogous to the specification of the rate-of-return constraint in most of the recent literature.

²⁹ It has been suggested to me that for some regulatory authorities this upper boundary is so high that many firms can comfortably try to maximize profits with prices fixed or declining without having to worry about the upper bound constraint's being binding. If firms should have to raise their prices, however, the lower bound constraint comes into play.

constraint, although it is not the nice clean one of the Averch-Johnson literature. Even so, we might expect firms operating on either the upper or lower boundary to exhibit Averch-Johnson type input distortions, although as long as capital is not a regressive input, firms operating at the higher rate-of-return boundary will exhibit less of a K/L distortion than those operating on the lower boundary *ceteris paribus*. When we introduce inflation, regulatory lag, and pricing based on historical average costs, we might expect all firms to move toward cost minimization. The high profit firms behave in this way in order to maintain their profits in the face of fixed output prices. The low profit firms exhibit this behavior in an effort to earn the rate of return they have been allowed, when prices have been set to yield this rate of return based on historical average costs which may be 18 months old when the new prices go into effect.

By and large the world of the regulated firm is far more complicated than much of the recent theoretical literature might lead one to believe. We have no doubt gained many valuable insights into the implications of rate-of-return regulation in the context of the neo-classical paradigm in which the regulatory process has been cast. It does appear, however, that our understanding of what is going on in the regulatory process is not so good that we should restrict ourselves to a one-paradigm approach. Particular paradigms not only lead to particular types of solutions, but may also restrict the kinds of questions that are asked.³⁰

In the general context of applied microeconomics, positive theoretical and empirical work aimed at constructing structural models of firm and organizational decision-making processes is not only intellectually satisfying, but should lead to a better understanding of firm and market behavior, better predictions of behavior and performance, and more useful insights for public policy analysis. Our ability to explain such things as technical change and innovation among firms and industries, product differentiation, merger activity, oligopoly pricing behavior, etc., is just not so good that continuing research into firm and organizational decision-making processes cannot yield useful insights.

The research effort presented here has attempted to move in this direction. Although useful insights into the pricing behavior of regulated firms have been gained, a framework for the analysis of price responsiveness to cost changes presented, and some new insights into efficiency incentives faced by regulated firms suggested, the work only scratches the surface. Questions of output and investment decisions have not even been mentioned. More importantly the questions of how decision rules change over time and why they change in certain directions for microeconomic units in general, are questions which the behavioral theory of the firm literature has only begun to tackle. As Scherer points out,³¹ the research effort necessary to answer these questions may be very different from that carried on in the past, but the returns from a successful search may be very high.

³⁰ This is a general proposition in some of the philosophy of science literature. See Kuhn [11].

³¹ Scherer [14].

TABLE 1

PROBIT INCREASE MODEL (1a)

OBSERVATION	I	F(I)%	INC*	OBSERVATION	I	F(I)%	INC*
1	1 18	88 10	1	33	-3 92	0.00	0
2	**	**		34	-1.87	3 07	0
3	2.39	99.12	1	35	-0.65	25 78	1
4	**	**		36	**	**	
5	-1.46	7.21	0	37	-1.14	12.51	0
6	2.45	99 29	1	38	-1 71	4.36	0
7	**	**		39	-1 42	8 78	0
8	**	**		40	-2.74	0.31	0
9	4.68	100.00	1	41	-4.86	0.00	0
10	**	**		42	-2.66	0.39	0
11	-0.42	33.72	0	43	-2.17	1.50	0
12	-1 28	10 03	0	44	-0.49	31.21	0
13	-4.03	0.00	0	45	0 92	82.12	1
14	-3 53	0 05	0	46	**	**	
15	-2.11	1.84	0	47	-0.23	40 90	1
16	-3 89	0 00	0	48	-3.33	0.05	0
17	-4.39	0 00	0	49	0.38	64 80	0
18	-3 67	0.00	1	50	-1 24	10.75	0
19	-2.81	0.25	0	51	-1 66	4 85	0
20	-2.59	0 48	0	52	-0 61	27.09	0
21	-2 98	0.14	0	53	2.48	99.34	1
22	-3 22	0 05	0	54	**	**	
23	-3 63	0.00	0	55	-2.32	1.02	0
24	-3 18	0.08	0	56	-2.10	1.69	0
25	-2.87	0.21	0	57	-1 12	13.14	0
26	-2.23	1.29	0	58	-2 32	1.02	0
27	-1 51	6.55	1	59	-6.06	0.00	0
28	-4 11	0 00	0	60	-4.85	0.00	0
29	-4.61	0.00	0	61	-1.99	2.33	0
30	-2.37	0.89	0	62	-0.69	24 51	0
31	-2 12	1 70	0	63	-1.44	7.49	0
32	-3.94	0 00	0				

*INC = 1, RATE INCREASE REQUESTS
= 0, OTHERWISE.

**YEARS THAT A PREVIOUS RATE INCREASE REQUEST WAS PENDING.

Appendix 1

■ Tables 1, 2, and 3 present the calculated probabilities based on the three probit models and the actual behavior of the firms. In Table 1 the years where a case is pending have been left blank. For models (1a) and (1b) there is a fairly high correspondence between high predicted probabilities of a rate increase and the actual behavior of the firms. Several observations do stand out, however, in which the predicted probabilities of a rate increase were very low and yet the firms in question actually petitioned for a rate increase. Several of these observations are examined further to see if the observed behavior can be explained.

Observation (18) is for the Rochester Gas and Electric Company for the year 1969. The predicted probability of a rate increase request is 0.00 for model (1a) and only 1.22 percent for model (1b) and yet a rate increase was observed. This is in some sense a special case. Two forces apparently prompted this company to ask for a rate increase

Calculated probabilities and actual behavior

TABLE 2

PROBIT INCREASE· MODEL(1b)

OBSERVATION	I	F(I)%	INC*	OBSERVATION	I	F(I)%	INC*
1	1.12	86.86	1	33	-2.68	0.37	0
2	1.75	95.99	1	34	-1.21	11.31	0
3	2.05	97.98	1	35	-0.37	35.57	1
4	1.86	96.93	1	36	1.09	85.99	1
5	-0.89	18.67	0	37	-0.78	22.77	0
6	2.07	98.08	1	38	-1.28	9.85	0
7	1.72	95.93	1	39	-1.00	15.87	0
8	0.62	73.24	1	40	-2.33	0.99	0
9	3.61	99.99	1	41	-3.45	0.03	0
10	1.07	85.77	1	42	-2.09	1.83	0
11	-0.19	42.47	0	43	-1.63	5.16	0
12	-0.99	15.87	0	44	-0.19	42.47	0
13	-2.99	0.14	0	45	0.91	82.89	1
14	-2.34	0.96	0	46	0.25	59.87	1
15	-1.47	7.08	0	47	-0.49	30.85	1
16	-2.75	0.30	0	48	-2.25	1.22	0
17	-2.97	0.15	0	49	0.31	62.17	0
18	-2.25	1.22	1	50	-1.05	14.69	0
19	-1.87	3.07	0	51	-1.01	15.62	0
20	-1.94	2.62	0	52	-0.54	29.46	0
21	-2.17	1.50	0	53	2.19	98.57	1
22	-2.22	1.32	0	54	3.82	100.00	1
23	-2.48	0.68	0	55	-1.64	5.05	0
24	-2.15	1.58	0	56	-1.45	7.35	0
25	-1.95	2.56	0	57	-0.51	30.50	0
26	-1.49	6.81	0	58	-1.76	3.92	0
27	-0.72	23.58	1	59	-4.18	0.00	0
28	-3.14	0.08	0	60	3.12	0.07	0
29	-3.29	0.05	0	61	-0.87	18.94	0
30	-1.70	4.46	0	62	-0.10	46.02	0
31	-1.56	5.94	0	63	-0.35	36.32	0
32	-2.89	0.19	0				

*INC = 1, RATE INCREASE REQUEST MADE OR PENDING
= 0, OTHERWISE.

in 1969. The first factor was that its bond rating had just been lowered in the market because the firm's interest coverage was too low. As indicated previously, firms attempt to maintain their bond ratings. The discrepancy in this case is thus partially a result of our inability to construct an ideal measure of the interest coverage pressures being felt by the firm. The second factor is that the firm was about to put on line a very large generating facility amounting to nearly 50 percent of its plant in service. The firm wished to have its prices adjusted upward to take into account this new plant, which had been financed at capital costs substantially above its historic cost of capital.³²

For observation (27), model (1a) predicted that the probability of an increase request was 6.55 percent while model (1b) predicted the probability of an increase request was 23.58 percent, yet the firm petitioned for a rate increase. Observation 27 corresponds to the Long Island Lighting Company for the year 1969. This appears to be a case where the firm is exploiting its "good will" with the Com-

TABLE 3

PROBIT DECREASE MODEL(2)

OBSERVATION	I	F(I)%	DEC*	OBSERVATION	I	F(I)%	DEC*
1	-2.72	0.33	0	33	-0.23	40.40	0
2	-2.18	1.66	0	34	-1.25	10.56	0
3	-3.09	0.10	0	35	-2.03	2.12	0
4	-1.82	3.44	0	36	-2.42	0.78	0
5	-0.45	32.64	0	37	-1.01	15.62	0
6	-3.11	0.09	0	38	-0.77	22.66	-1
7	-2.12	1.70	0	39	-1.14	12.71	-1
8	-2.07	1.92	0	40	0.12	54.78	-1
9	-4.36	0.00	0	41	0.01	50.40	0
10	-2.64	0.41	0	42	-0.85	19.77	0
11	-1.89	2.94	0	43	-1.04	14.92	0
12	-0.54	29.10	0	44	-1.41	7.93	0
13	0.85	80.23	-1	45	-2.05	2.02	0
14	-0.17	43.25	0	46	-2.19	1.43	0
15	-0.83	20.33	-1	47	-0.22	41.29	0
16	0.08	53.19	0	48	0.13	55.17	-1
17	0.61	72.91	0	49	-1.80	3.59	0
18	0.21	58.32	0	50	-0.48	31.56	0
19	-0.28	38.97	0	51	-1.50	6.68	0
20	-0.08	53.19	0	52	-1.40	8.08	0
21	0.03	51.20	-1	53	-4.13	0.00	0
22	0.17	56.75	-1	54	-3.99	0.00	0
23	0.46	67.72	-1	55	-0.35	36.32	0
24	-0.09	46.39	0	56	-0.11	45.62	-1
25	-0.30	38.21	-1	57	-0.81	20.90	-1
26	-0.28	38.97	0	58	0.41	65.91	-1
27	-1.20	11.51	0	59	1.93	97.32	-1
28	0.54	70.54	0	60	0.98	83.65	-1
29	0.36	64.06	-1	61	-0.84	20.05	0
30	-0.40	34.46	0	62	-0.64	26.11	0
31	-0.62	26.76	-1	63	-0.56	28.77	0
32	-0.60	72.57	-1				

*DEC = -1, RATE DECREASE PROPOSAL
= 0, OTHERWISE

mission in an attempt to get a rate increase approved at a higher level of recent earnings than would a firm with a smaller amount of "good will." LILCO had built up a high level of good will with the Commission during the 1960's through several rate decreases and generally efficient performance. In filing for a rate increase in 1969, LILCO most likely expected more favorable treatment from the Commission than would a firm with a smaller stock of good will.

Observations (35) (N.Y. State Electric and Gas, 1968) and (47) (Niagara Mohawk, 1962) also exhibit fairly low predicted probabilities of a rate increase for both models and yet increase requests are observed. In both cases a *negative* growth rate in earnings per share is observed for either the current or previous year. This indicates that an actual decline in earnings per share may prompt a rate increase request more readily than will positive growth rates in earnings per share yielding the same decision index.

Observation (49) (Niagara Mohawk, 1964) yields a predicted probability of 64.80 percent for model (1a) and 62.17 percent for model (1b) and yet a rate increase request did not take place. In

case 21439³³ (August 1961) the Commission allowed Niagara Mohawk a price increase considerably less than requested because the Commission maintained that the company was entering a transitional period during which it would be both obtaining cheaper power from the Niagara Power Authority³⁴ and phasing out its own high cost equipment. Basically, the Commission held that the test year data did not properly reflect the costs the company would be facing over the next few years. Niagara Mohawk may not have made a request for an increase in 1964, fearing that the Commission would reject it pending more experience with the lower cost Niagara Power Authority electricity. As it turned out, Niagara Mohawk's rate of return on electric operations increased dramatically after 1964.

Turning to Table 3 we observe a fairly good correspondence between predicted probabilities of a rate decrease and actual decrease behavior. There are several instances in which there is a high predicted probability of a rate decrease and yet there is no decrease. This situation appears, however, to occur mainly for firms that have a series of rate decreases in the sample, note in particular observations (16), (17), and (18) where the company had filed rate decreases twice in the previous few years and observations (25) and (28) where the relevant companies made a series of rate decreases during the nine-year sample period. Such firms may have kept the high profits for a given year based on their "good record" in passing on previous years' profits.

We also observe cases where the predicted probability of a rate decrease is low and yet decreases are observed. Note especially observations (38) and (39). This behavior might be explained as attempts by the firms to build up good will with the Commission.

Appendix 2

Regulatory lag and rate increases

■ It can take anywhere from six months to a year, and sometimes longer, from the day a firm files a rate increase request to the time litigation before the Commission is completed and a final decision is reached. During this period of time the firm must continue to provide service at prevailing rates (although in one case in the sample the Commission used its power to allow a temporary rate increase). We are interested in determining whether or not firms continue to experience poor financial performance during this lag period and therefore suffer short term losses, due to hearing lag time, for which they may never be compensated.

Table 4 includes those observations for which a case was in progress at the time of the observation. The question is whether financial indicators would have provided a strong impetus for a rate increase if a case had *not* been in progress at the time. This can be investigated by calculating the probabilities of a rate increase in model (1a) for those years which were omitted because a case was pending. We are asking the question: What was the probability of a rate increase in the lag years, given the exogenous variables, and assuming

³³ 40 PUR 3d, p. 429.

³⁴ The Niagara Power Authority produces electricity from facilities near Niagara Falls and sells it at wholesale to companies in New York State.

TABLE 4

REGULATORY LAG TABLE. CALCULATED PROBABILITIES OF A RATE INCREASE REQUEST FOR LAG YEARS

OBSERVATION	I_{INC}^{1-A}	I_{INC}^{1-B}	$F(I_{INC})\%$	$F(I^*_{INC})\%$
2	2.80	1.75	99.74	95.99
4	3.12	1.86	99.91	96.93
7	2.60	1.72	99.53*	95.93*
8	0.49	0.62	68.79*	73.24*
10	1.87	1.07	96.93	85.77
36	1.39	1.09	91.77	85.99
46	0.19	0.25	57.53	59.87
54	5.80	3.82	100.00	100.00

*TEMPORARY RATE INCREASES WERE ALLOWED PENDING THE OUTCOME OF THE HEARINGS. THIS CASE TOOK AN INORDINATELY LONG TIME CON ED FILED ITS INITIAL RATE INCREASE IN OCTOBER 1966 AND LITIGATION WAS NOT COMPLETED UNTIL JUNE 27, 1968.

that a case was *not* pending in those years? We can also look at the corresponding values for model (1b). (See Table 4.)

If financial indicators and the approval variable would have indicated a rate increase petition in the absence of a pending case, the probabilities of a rate increase for the lag years, using model (1a), should be very high. Similarly, model (1b) should produce a high predicted probability for those observations. If temporary rate increases were allowed, the financial plight of the companies involved should have been alleviated somewhat and there should be a decline in the probability of a rate increase request from the year the request was made to the lag year.

The high calculated probabilities in the last two columns of Table 4 indicate that firms that found it necessary to file for a rate increase in one period continued to experience financial difficulty during the period in which their cases were being decided. These values indicate that if the performance during the time the case was pending had been observed in a year when a case was not pending, such performance would have led the firm to request a rate increase. This is not at all surprising since the decision by the firm to file for an increase involves some judgment that financial difficulties at current price levels are not merely the result of one period of poor performance but rather will continue in the future. In addition, in the case where a temporary rate increase was allowed, a decline in the probabilities of an increase request from the request year to the lag year is observed for both models. Temporary rate increases appear to alleviate some of the financial pressure on firms awaiting a decision on their request.

References

1. ALLISON, G. *Essence of Decision: Explaining the Cuban Missile Crises* New York: Little, Brown, 1972.

2. AVERCH, H. AND JOHNSON, L. "Behavior of the Firm under Regulatory Constraint," *The American Economic Review*, Vol. 52, No. 5 (December 1962), pp. 1053-1069.

3. BAUMOL, W. AND KLEVORICK, A. "Input Choices and Rate of Return Regulation: An Overview of the Discussion." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 162-190.

4. CYERT, T. AND MARCH, J. *A Behavioral Theory of the Firm* Englewood Cliffs, N. J.: Prentice Hall, 1963.
5. FRIEDMAN, M. "The Methodology of Positive Economics," in *Essays in Positive Economics*, University of Chicago Press, 1953.
6. GOLDBERGER, A. *Econometric Theory*. New York: John Wiley & Sons, 1964.
7. JOSKOW, P. "A Behavioral Theory of Public Utility Regulation." Unpublished Ph.D. dissertation, Yale University, 1972.
8. ———. "The Determination of the Allowed Rate of Return in a Formal Regulatory Proceeding." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 2 (Autumn 1972), pp. 632–644.
9. KAHN, A. "The Graduated Fair Return: Comment." *The American Economic Review*, Vol. 58, No. 1 (March 1968), pp. 170–173.
10. KLEVORICK, A. "The 'Optimal' Fair Rate of Return." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 122–153.
11. KUHN, T. *The Structure of Scientific Revolutions*. 2nd ed. Chicago: University of Chicago Press, 1970.
12. LITZENBERGER, R. AND RAO, C. "Estimates of the Marginal Rate of Time Preference and Average Risk Aversion of Investors in Electric Utility Shares: 1960–1966." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 265–277.
13. MOOD, A. *Introduction to the Theory of Statistics*. New York: McGraw-Hill, 1950.
14. SCHERER, F. *Industrial Market Structure and Economic Performance*. Chicago: Rand McNally & Co., 1970.
15. SIMON, H. "Theories of Decision-Making in Economics and Behavioral Science." *The American Economic Review*, Vol. 49, No. 3 (June 1959), pp. 253–283.
16. TOBIN, J. "An Application of Multivariate Probit Analysis to Economic Survey Data." Cowles Foundation Discussion Paper No. 1, 1955.
17. WOOD, STRUTHERS AND WINTHROP, INCORPORATED. "The Changing Pattern of Utility Rates of Return." Mimeographed. September 1971.

Theory of rational option pricing

Robert C. Merton

Assistant Professor of Finance
Sloan School of Management
Massachusetts Institute of Technology

The long history of the theory of option pricing began in 1900 when the French mathematician Louis Bachelier deduced an option pricing formula based on the assumption that stock prices follow a Brownian motion with zero drift. Since that time, numerous researchers have contributed to the theory. The present paper begins by deducing a set of restrictions on option pricing formulas from the assumption that investors prefer more to less. These restrictions are necessary conditions for a formula to be consistent with a rational pricing theory. Attention is given to the problems created when dividends are paid on the underlying common stock and when the terms of the option contract can be changed explicitly by a change in exercise price or implicitly by a shift in the investment or capital structure policy of the firm. Since the deduced restrictions are not sufficient to uniquely determine an option pricing formula, additional assumptions are introduced to examine and extend the seminal Black-Scholes theory of option pricing. Explicit formulas for pricing both call and put options as well as for warrants and the new "down-and-out" option are derived. The effects of dividends and call provisions on the warrant price are examined. The possibilities for further extension of the theory to the pricing of corporate liabilities are discussed.

■ The theory of warrant and option pricing has been studied extensively in both the academic and trade literature.¹ The approaches taken range from sophisticated general equilibrium models to ad hoc statistical fits. Because options are specialized and relatively unimportant financial securities, the amount of time and space devoted to the development of a pricing theory might be questioned. One justification is that, since the option is a particularly simple type of contingent-claim asset, a theory of option pricing may lead to a general theory of contingent-claims pricing. Some have argued that all such securities can be expressed as combinations of basic option contracts, and, as such, a theory of option pricing constitutes a

1. Introduction

Robert C. Merton received the B.S. in engineering mathematics from Columbia University's School of Engineering and Applied Science (1966), the M.S. in applied mathematics from the California Institute of Technology (1967), and the Ph.D. from the Massachusetts Institute of Technology (1970). Currently he is Assistant Professor of Finance at M.I.T., where he is conducting research in capital theory under uncertainty.

The paper is a substantial revision of sections of Merton [34] and [29]. I am particularly grateful to Myron Scholes for reading an earlier draft and for his comments. I have benefited from discussion with P. A. Samuelson and F. Black. I thank Robert K. Merton for editorial assistance. Any errors remaining are mine. Aid from the National Science Foundation is gratefully acknowledged.

¹ See the bibliography for a substantial, but partial, listing of papers.

theory of contingent-claims pricing.² Hence, the development of an option pricing theory is, at least, an intermediate step toward a unified theory to answer questions about the pricing of a firm's liabilities, the term and risk structure of interest rates, and the theory of speculative markets. Further, there exist large quantities of data for testing the option pricing theory.

The first part of the paper concentrates on laying the foundations for a rational theory of option pricing. It is an attempt to derive theorems about the properties of option prices based on assumptions sufficiently weak to gain universal support. To the extent it is successful, the resulting theorems become necessary conditions to be satisfied by any rational option pricing theory.

As one might expect, assumptions weak enough to be accepted by all are not sufficient to determine uniquely a rational theory of option pricing. To do so, more structure must be added to the problem through additional assumptions at the expense of losing some agreement. The Black and Scholes (henceforth, referred to as B-S) formulation³ is a significant "break-through" in attacking the option problem. The second part of the paper examines their model in detail. An alternative derivation of their formula shows that it is valid under weaker assumptions than they postulate. Several extensions to their theory are derived.

2. Restrictions on rational option pricing⁴

■ An "American"-type warrant is a security, issued by a company, giving its owner the right to purchase a share of stock at a given ("exercise") price on or before a given date. An "American"-type call option has the same terms as the warrant except that it is issued by an individual instead of a company. An "American"-type put option gives its owner the right to sell a share of stock at a given exercise price on or before a given date. A "European"-type option has the same terms as its "American" counterpart except that it cannot be surrendered ("exercised") before the last date of the contract. Samuelson⁵ has demonstrated that the two types of contracts may not have the same value. All the contracts may differ with respect to other provisions such as antidilution clauses, exercise price changes, etc. Other option contracts such as strips, straps, and straddles, are combinations of put and call options.

The principal difference between valuing the call option and the warrant is that the aggregate supply of call options is zero, while the aggregate supply of warrants is generally positive. The "bucket shop" or "incipient" assumption of zero aggregate supply⁶ is useful

² See Black and Scholes [4] and Merton [29].

³ In [4].

⁴ This section is based on Merton [34] cited in Samuelson and Merton [43], p. 43, footnote 6.

⁵ In [42].

⁶ See Samuelson and Merton [43], p. 26 for a discussion of "incipient" analysis. Essentially, the incipient price is such that a slightly higher price would induce a positive supply. In this context, the term "bucket shop" was coined in oral conversation by Paul Samuelson and is based on the (now illegal) 1920's practice of side-bets on the stock market.

Myron Scholes has pointed out that if a company sells a warrant against stock already *outstanding* (not just authorized), then the incipient analysis is valid as well. (E.g., Amerada Hess selling warrants against shares of Louisiana Land

because the probability distribution of the stock price return is unaffected by the creation of these options, which is not in general the case when they are issued by firms in positive amounts.⁷ The “bucket-shop” assumption is made throughout the paper although many of the results derived hold independently of this assumption.

The notation used throughout is: $F(S, \tau; E)$ — the value of an American warrant with exercise price E and τ years before expiration, when the price per share of the common stock is S ; $f(S, \tau; E)$ — the value of its European counterpart; $G(S, \tau; E)$ — the value of an American put option; and $g(S, \tau; E)$ — the value of its European counterpart.

From the definition of a warrant and limited liability, we have that

$$F(S, \tau; E) \geq 0; \quad f(S, \tau; E) \geq 0 \quad (1)$$

and when $\tau = 0$, at expiration, both contracts must satisfy

$$F(S, 0; E) = f(S, 0; E) = \text{Max}[0, S - E]. \quad (2)$$

Further, it follows from conditions of arbitrage that

$$F(S, \tau; E) \geq \text{Max}[0, S - E]. \quad (3)$$

In general, a relation like (3) need not hold for a European warrant.

Definition: Security (portfolio) A is *dominant* over security (portfolio) B , if on some known date in the future, the return on A will exceed the return on B for some possible states of the world, and will be at least as large as on B , in all possible states of the world.

Note that in perfect markets with no transactions costs and the ability to borrow and short-sell without restriction, the existence of a dominated security would be equivalent to the existence of an arbitrage situation. However, it is possible to have dominated securities exist without arbitrage in imperfect markets. If one assumes something like “symmetric market rationality” and assumes further that investors prefer more wealth to less,⁸ then any investor willing to purchase security B would prefer to purchase A .

Assumption 1: A necessary condition for a rational option pricing theory is that the option be priced such that it is neither a dominant nor a dominated security.

Given two American warrants on the same stock and with the same exercise price, it follows from Assumption 1, that

$$F(S, \tau_1; E) \geq F(S, \tau_2; E) \quad \text{if} \quad \tau_1 > \tau_2, \quad (4)$$

and that

$$F(S, \tau; E) \geq f(S, \tau; E). \quad (5)$$

Further, two warrants, identical in every way except that one has a larger exercise price than the other, must satisfy

$$\begin{aligned} F(S, \tau; E_1) &\leq F(S, \tau; E_2) \\ f(S, \tau; E_1) &\leq f(S, \tau; E_2) \quad \text{if} \quad E_1 > E_2. \end{aligned} \quad (6)$$

and Exploration stock it owns and City Investing selling warrants against shares of General Development Corporation stock it owns.)

⁷ See Merton [29], Section 2.

⁸ See Modigliani and Miller [35], p. 427, for a definition of “symmetric market rationality.”

Because the common stock is equivalent to a perpetual ($\tau = \infty$) American warrant with a zero exercise price ($E = 0$), it follows from (4) and (6) that

$$S \geq F(S, \tau; E), \quad (7)$$

and from (1) and (7), the warrant must be worthless if the stock is, i.e.,

$$F(0, \tau; E) = f(0, \tau; E) = 0. \quad (8)$$

Let $P(\tau)$ be the price of a riskless (in terms of default), discounted loan (or "bond") which pays one dollar, τ years from now. If it is assumed that current and future interest rates are positive, then

$$1 = P(0) > P(\tau_1) > P(\tau_2) > \cdots > P(\tau_n) \\ \text{for } 0 < \tau_1 < \tau_2 < \cdots < \tau_n, \quad (9)$$

at a given point in calendar time.

Theorem 1. If the exercise price of a European warrant is E and if no payouts (e.g. dividends) are made to the common stock over the life of the warrant (or alternatively, if the warrant is protected against such payments), then $f(S, \tau; E) \geq \text{Max}[0, S - EP(\tau)]$.

Proof: Consider the following two investments:

- A: Purchase the warrant for $f(S, \tau; E)$;
Purchase E bonds at price $P(\tau)$ per bond.
Total investment: $f(S, \tau; E) + EP(\tau)$.
- B: Purchase the common stock for S .
Total investment: S .

Suppose at the end of τ years, the common stock has value S^* . Then, the value of B will be S^* . If $S^* \leq E$, then the warrant is worthless and the value of A will be $0 + E = E$. If $S^* > E$, then the value of A will be $(S^* - E) + E = S^*$. Therefore, unless the current value of A is at least as large as B , A will dominate B . Hence, by Assumption 1, $f(S, \tau; E) + EP(\tau) \geq S$, which together with (1), implies that $f(S, \tau; E) \geq \text{Max}[0, S - EP(\tau)]$. Q.E.D.

From (5), it follows directly that Theorem 1 holds for American warrants with a fixed exercise price over the life of the contract. The right to exercise an option prior to the expiration date always has nonnegative value. It is important to know when this right has zero value, since in that case, the values of an European and American option are the same. In practice, almost all options are of the American type while it is always easier to solve analytically for the value of an European option. Theorem 1 significantly tightens the bounds for rational warrant prices over (3). In addition, it leads to the following two theorems.

Theorem 2. If the hypothesized conditions for Theorem 1 hold, an American warrant will never be exercised prior to expiration, and hence, it has the same value as a European warrant.

Proof: If the warrant is exercised, its value will be $\text{Max}[0, S - E]$. But from Theorem 1, $F(S, \tau; E) \geq \text{Max}[0, S - EP(\tau)]$, which is larger than $\text{Max}[0, S - E]$ for $\tau > 0$ because, from (9), $P(\tau) < 1$. Hence, the warrant is always worth more "alive" than "dead." Q.E.D.

Theorem 2 suggests that if there is a difference between the American and European warrant prices which implies a positive probability of a premature exercise, it must be due to unfavorable changes in the exercise price or to lack of protection against payouts to the common stocks. This result is consistent with the findings of Samuelson and Merton.⁹

It is a common practice to refer to $\text{Max}[0, S - E]$ as the *intrinsic value* of the warrant and to state that the warrant must always sell for at least its intrinsic value [condition (3)]. In light of Theorems 1 and 2, it makes more sense to define $\text{Max}[0, S - EP(\tau)]$ as the intrinsic value. The latter definition reflects the fact that the amount of the exercise price need not be paid until the expiration date, and $EP(\tau)$ is just the present value of that payment. The difference between the two values can be large, particularly for long-lived warrants, as the following theorem demonstrates.

Theorem 3. If the hypothesized conditions for Theorem 1 hold, the value of a perpetual ($\tau = \infty$) warrant must equal the value of the common stock.

Proof: From Theorem 1, $F(S, \infty; E) \geq \text{Max}[0, S - EP(\infty)]$. But, $P(\infty) = 0$, since, for positive interest rates, the value of a discounted loan payable at infinity is zero. Therefore, $F(S, \infty; E) \geq S$. But from (7), $S \geq F(S, \infty; E)$. Hence, $F(S, \infty; E) = S$. Q.E.D.

Samuelson, Samuelson and Merton, and Black and Scholes¹⁰ have shown that the price of a perpetual warrant equals the price of the common stock for their particular models. Theorem 3 demonstrates that it holds independent of any stock price distribution or risk-averse behavioral assumptions.¹¹

The inequality of Theorem 1 demonstrates that a finite-lived, rationally-determined warrant price must be a function of $P(\tau)$. For if it were not, then, for some sufficiently small $P(\tau)$ (i.e., large interest rate), the inequality of Theorem 1 would be violated. From the form of the inequality and previous discussion, this direct dependence on the interest rate seems to be "induced" by using as a variable, the exercise price instead of the present value of the exercise price (i.e., I conjecture that the pricing function, $F[S, \tau; E, P(\tau)]$, can be written as $W(S, \tau; e)$, where $e = EP(\tau)$.¹² If this is so, then the qualitative effect of a change in P on the warrant price would be similar to a change in the exercise price, which, from (6), is negative. Therefore, the warrant price should be an increasing function of the interest rate. This finding is consistent with the theoretical models of Samuel-

⁹ In [43], p. 29 and Appendix 2.

¹⁰ In [42], [43], and [4], respectively.

¹¹ It is a bit of a paradox that a perpetual warrant with a positive exercise price should sell for the same price as the common stock (a "perpetual warrant" with a zero exercise price), and, in fact, the few such outstanding warrants do not sell for this price. However, it must be remembered that one assumption for the theorem to obtain is that no payouts to the common stock will be made over the life of the contract which is almost never true in practice. See Samuelson and Merton [43], pp. 30-31, for further discussion of the paradox.

¹² The only case where the warrant price does not depend on the exercise price is the perpetuity, and the only case where the warrant price does not depend on $P(\tau)$ is when the exercise price is zero. Note that in both cases, $e = 0$, (the former because $P(\infty) = 0$, and the latter because $E = 0$), which is consistent with our conjecture.

son and Merton and Black and Scholes and with the empirical study by Van Horne.¹³

Another argument for the reasonableness of this result comes from recognizing that a European warrant is equivalent to a long position in the common stock levered by a limited-liability, discount loan, where the borrower promises to pay E dollars at the end of τ periods, but in the event of default, is only liable to the extent of the value of the common stock at that time.¹⁴ If the present value of such a loan is a decreasing function of the interest rate, then, for a given stock price, the warrant price will be an increasing function of the interest rate.

We now establish two theorems about the effect of a change in exercise price on the price of the warrant.

Theorem 4. If $F(S, \tau; E)$ is a rationally determined warrant price, then F is a convex function of its exercise price, E .

Proof: To prove convexity, we must show that if

$$E_3 \equiv \lambda E_1 + (1 - \lambda)E_2,$$

then for every λ , $0 \leq \lambda \leq 1$,

$$F(S, \tau; E_3) \leq \lambda F(S, \tau; E_1) + (1 - \lambda)F(S, \tau; E_2).$$

We do so by a dominance argument similar to the proof of Theorem 1. Let portfolio A contain λ warrants with exercise price E_1 and $(1 - \lambda)$ warrants with exercise price E_2 where by convention, $E_2 > E_1$. Let portfolio B contain one warrant with exercise price E_3 . If S^* is the stock price on the date of expiration, then by the convexity of $\text{Max}[0, S^* - E]$, the value of portfolio A ,

$$\lambda \text{Max}[0, S^* - E_1] + (1 - \lambda) \text{Max}[0, S^* - E_2],$$

will be greater than or equal to the value of portfolio B ,

$$\text{Max}[0, S^* - \lambda E_1 - (1 - \lambda)E_2].$$

Hence, to avoid dominance, the current value of portfolio B must be less than or equal to the current value of portfolio A . Thus, the theorem is proved for a European warrant. Since nowhere in the argument is any factor involving τ used, the same results would obtain if the warrants in the two portfolios were exercised prematurely. Hence, the theorem holds for American warrants. Q.E.D.

Theorem 5. If $f(S, \tau; E)$ is a rationally determined European warrant price, then for $E_1 < E_2$, $-P(\tau)(E_2 - E_1) \leq f(S, \tau; E_2) - f(S, \tau; E_1) \leq 0$. Further, if f is a differentiable function of its exercise price, $-P(\tau) \leq \partial f(S, \tau; E)/\partial E \leq 0$.

Proof: The right-hand inequality follows directly from (6). The left-hand inequality follows from a dominance argument. Let portfolio A contain a warrant to purchase the stock at E_2 and $(E_2 - E_1)$ bonds at price $P(\tau)$ per bond. Let portfolio B contain a warrant to purchase the stock at E_1 . If S^* is the stock price on the

¹³ In [43], [4], and [54], respectively.

¹⁴ Stiglitz [51], p. 788, introduces this same type loan as a sufficient condition for the Modigliani-Miller Theorem to obtain when there is a positive probability of bankruptcy.

date of expiration, then the terminal value of portfolio A ,

$$\text{Max}[0, S^* - E_2] + (E_2 - E_1),$$

will be greater than the terminal value of portfolio B , $\text{Max}[0, S^* - E_1]$, when $S^* < E_2$, and equal to it when $S^* \geq E_2$. So, to avoid dominance, $f(S, \tau; E_1) \leq f(S, \tau; E_2) + P(\tau)(E_2 - E_1)$. The inequality on the derivative follows by dividing the discrete-change inequalities by $(E_2 - E_1)$ and taking the limit as E_2 tends to E_1 . Q.E.D.

If the hypothesized conditions for Theorem 1 hold, then the inequalities of Theorem 5 hold for American warrants. Otherwise, we only have the weaker inequalities, $-(E_2 - E_1) \leq F(S, \tau; E_2) - F(S, \tau; E_1) \leq 0$ and $-1 \leq \partial F(S, \tau; E)/\partial E \leq 0$.

Let $Q(t)$ be the price per share on a common stock at time t and $F_Q(Q, \tau; E_Q)$ be the price of a warrant to purchase one share of stock at price E_Q on or before a given date τ years in the future, when the current price of the common stock is Q .

Theorem 6. If k is a positive constant; $Q(t) = kS(t)$; $E_Q = kE$, then $F_Q(Q, \tau; E_Q) \equiv kF(S, \tau; E)$ for all S, τ, E and each k .

Proof: Let S^* be the value of the common stock with initial value S when both warrants either are exercised or expire. Then, by the hypothesized conditions of the theorem, $Q = Q^* \equiv kS^*$ and $E_Q = kE$. The value of the warrant on Q will be $\text{Max}[0, Q^* - E_Q] = k \text{Max}[0, S^* - E]$ which is k times the value of the warrant on S . Hence, to avoid dominance of one over the other, the value of the warrant on Q must sell for exactly k times the value of the warrant on S . Q.E.D.

The implications of Theorem 6 for restrictions on rational warrant pricing depend on what assumptions are required to produce the hypothesized conditions of the theorem. In its weakest form, it is a dimensional theorem where k is the proportionality factor between two units of account (e.g., $k = 100$ cents/dollar). If the stock and warrant markets are purely competitive, then it can be interpreted as a scale theorem. Namely, if there are no economies of scale with respect to transactions costs and no problems with indivisibilities, then k shares of stock will always sell for exactly k times the value of one share of stock. Under these conditions, the theorem states that a warrant to buy k shares of stock for a total of (kE) dollars when the stock price per share is S dollars, is equal in value to k times the price of a warrant to buy one share of the stock for E dollars, all other terms the same. Thus, the rational warrant pricing function is homogeneous of degree one in S and E with respect to scale, which reflects the usual constant returns to scale results of competition.

Hence, one can always work in standardized units of $E = 1$ where the stock price and warrant price are quoted in units of exercise price by choosing $k = 1/E$. Not only does this change of units eliminate a variable from the problem, but it is also a useful operation to perform prior to making empirical comparisons across different warrants where the dollar amounts may be of considerably different magnitudes.

Let $F_i(S_i, \tau_i; E_i)$ be the value of a warrant on the common stock of firm i with current price per share S_i when τ_i is the time to expiration and E_i is the exercise price.

Assumption 2. If $S_i = S_j = S$; $\tau_i = \tau_j = \tau$; $E_i = E_j = E$, and the returns per dollar on the stocks i and j are identically distributed, then $F_i(S, \tau; E) = F_j(S, \tau; E)$.

Assumption 2 implies that, from the point of view of the warrant holder, the only identifying feature of the common stock is its (*ex ante*) distribution of returns.

Define $z(t)$ to be the one-period random variable return per dollar invested in the common stock in period t . Let $Z(\tau) \equiv \prod_{t=1}^{\tau} z(t)$ be the τ -period return per dollar.

Theorem 7. If $S_i = S_j = S$, $i, j = 1, 2, \dots, n$;

$$Z_{n+1}(\tau) \equiv \sum_{i=1}^n \lambda_i Z_i(\tau)$$

for $\lambda_i \in [0, 1]$ and $\sum_{i=1}^n \lambda_i = 1$, then

$$F_{n+1}(S, \tau; E) \leq \sum_{i=1}^n \lambda_i F_i(S, \tau; E).$$

Proof: By construction, one share of the $(n+1)$ st security contains λ_i shares of the common stock of firm i , and by hypothesis, the price per share, $S_{n+1} = \sum_{i=1}^n \lambda_i S_i = S \sum_{i=1}^n \lambda_i = S$. The proof follows from a dominance argument. Let portfolio A contain λ_i warrants on the common stock of firm i , $i = 1, 2, \dots, n$. Let portfolio B contain one warrant on the $(n+1)$ st security. Let S_i^* denote the price per share on the common stock of the i th firm, on the date of expiration, $i = 1, 2, \dots, n$. By definition, $S_{n+1}^* = \sum_{i=1}^n \lambda_i S_i^*$. On the expiration date, the value of portfolio A , $\sum_{i=1}^n \lambda_i \text{Max}[0, S_i^* - E]$, is greater than or equal to the value of portfolio B , $\text{Max}[0, \sum_{i=1}^n \lambda_i S_i^* - E]$, by the convexity of $\text{Max}[0, S - E]$. Hence, to avoid dominance,

$$F_{n+1}(S, \tau; E) \leq \sum_{i=1}^n \lambda_i F_i(S, \tau; E). \text{ Q.E.D.}$$

Loosely, Theorem 7 states that a warrant on a portfolio is less valuable than a portfolio of warrants. Thus, from the point of view of warrant value, diversification “hurts,” as the following special case of Theorem 7 demonstrates:

Corollary. If the hypothesized conditions of Theorem 7 hold and if, in addition, the $z_i(t)$ are identically distributed, then

$$F_{n+1}(S, \tau; E) \leq F_i(S, \tau; E)$$

for $i = 1, 2, \dots, n$.

Proof: From Theorem 7, $F_{n+1}(S, \tau; E) \leq \sum_{i=1}^n \lambda_i F_i(S, \tau; E)$. By hypothesis, the $z_i(t)$ are identically distributed, and hence, so are the $Z_i(\tau)$. Therefore, by Assumption 2, $F_i(S, \tau; E) = F_j(S, \tau; E)$ for $i, j = 1, 2, \dots, n$. Since $\sum_{i=1}^n \lambda_i = 1$, it then follows that $F_{n+1}(S, \tau; E) \leq F_i(S, \tau; E)$, $i = 1, 2, \dots, n$. Q.E.D.

Theorem 7 and its Corollary suggest the more general proposition that the more risky the common stock, the more valuable the warrant. In order to prove the proposition, one must establish a careful definition of “riskiness” or “volatility.”

Definition: Security one is *more risky* than security two if $Z_1(\tau) = Z_2(\tau) + \epsilon$ where ϵ is a random variable with the property

$$E[\epsilon | Z_2(\tau)] = 0.$$

This definition of more risky is essentially one of the three (equivalent) definitions used by Rothschild and Stiglitz.¹⁵

Theorem 8. The rationally determined warrant price is a non-decreasing function of the riskiness of its associated common stock.

Proof: Let $Z(\tau)$ be the τ -period return on a common stock with warrant price, $F_Z(S, \tau; E)$. Let $Z_i(\tau) = Z(\tau) + \epsilon_i$, $i = 1, \dots, n$, where the ϵ_i are independently and identically distributed random variables satisfying $E[\epsilon_i | Z(\tau)] = 0$. By definition, security i is more risky than security Z , for $i = 1, \dots, n$. Define the random variable

$$\text{return } Z_{n+1}(\tau) \equiv \frac{1}{n} \sum_{i=1}^n Z_i(\tau) = Z(\tau) + \frac{1}{n} \sum_{i=1}^n \epsilon_i. \text{ Note that, by con-}$$

struction, the $Z_i(\tau)$ are identically distributed. Hence, by the Corollary to Theorem 7 with $\lambda_i = 1/n$, $F_{n+1}(S, \tau; E) \leq F_i(S, \tau; E)$ for $i = 1, 2, \dots, n$. By the law of large numbers, $Z_{n+1}(\tau)$ converges in probability to $Z(\tau)$ as $n \rightarrow \infty$, and hence, by Assumption 2, limit $F_{n+1}(S, \tau; E) = F_Z(S, \tau; E)$. Therefore, $F_Z(S, \tau; E) \leq F_i(S, \tau; E)$ for $i = 1, 2, \dots, n$. Q.E.D.

Thus, the more uncertain one is about the outcomes on the common stock, the more valuable is the warrant. This finding is consistent with the empirical study by Van Horne.¹⁶

To this point in the paper, no assumptions have been made about the properties of the distribution of returns on the common stock. If it is assumed that the $\{z(t)\}$ are independently distributed,¹⁷ then the distribution of the returns per dollar invested in the stock is independent of the initial level of the stock price, and we have the following theorem:

Theorem 9. If the distribution of the returns per dollar invested in the common stock is independent of the level of the stock price, then $F(S, \tau; E)$ is homogeneous of degree one in the stock price per share and exercise price.

Proof: Let $z_i(t)$ be the return per dollar if the initial stock price is S_i , $i = 1, 2$. Define $k = (S_2/S_1)$ and $E_2 = kE_1$. Then, by Theorem 6, $F_2(S_2, \tau; E_2) \equiv kF_2(S_1, \tau; E_1)$. By hypothesis, $z_1(t)$ and $z_2(t)$ are identically distributed. Hence, by Assumption 2, $F_2(S_1, \tau; E_1) = F_1(S_1, \tau; E_1)$. Therefore, $F_2(kS_1, \tau; kE_1) \equiv kF_1(S_1, \tau; E_1)$ and the theorem is proved. Q.E.D.

Although similar in a formal sense, Theorem 9 is considerably stronger than Theorem 6, in terms of restrictions on the warrant pricing function. Namely, given the hypothesized conditions of Theorem 9, one would expect to find in a table of rational warrant values for a given maturity, that the value of a warrant with exercise price E when the common stock is at S will be exactly k times as

¹⁵ The two other equivalent definitions are: (1) every risk averter prefers X to Y (i.e., $EU(X) \geq EU(Y)$, for all concave U); (2) Y has more weight in the tails than X . In addition, they show that if Y has greater variance than X , then it need not be more risky in the sense of the other three definitions. It should also be noted that it is the *total* risk, and not the *systematic* or portfolio risk, of the common stock which is important to warrant pricing. In [39], p. 225.

¹⁶ In [54].

¹⁷ Cf. Samuelson [42].

valuable as a warrant on the same stock with exercise price E/k when the common stock is selling for S/k . In general, this result will not obtain if the distribution of returns depends on the level of the stock price as is shown by a counter example in Appendix 1.

Theorem 10. If the distribution of the returns per dollar invested in the common stock is independent of the level of the stock price, then $F(S, \tau; E)$ is a convex function of the stock price.

Proof: To prove convexity, we must show that if

$$S_3 \equiv \lambda S_1 + (1 - \lambda)S_2,$$

then, for every λ , $0 \leq \lambda \leq 1$,

$$F(S_3, \tau; E) \leq \lambda F(S_1, \tau; E) + (1 - \lambda)F(S_2, \tau; E).$$

From Theorem 4,

$$F(1, \tau; E_3) \leq \gamma F(1, \tau; E_1) + (1 - \gamma)F(1, \tau; E_2),$$

for $0 \leq \gamma \leq 1$ and $E_3 = \gamma E_1 + (1 - \gamma)E_2$. Take $\gamma \equiv \lambda S_1/S_3$, $E_1 \equiv E/S_1$, and $E_2 \equiv E/S_2$. Multiplying both sides of the inequality by S_3 , we have that

$$S_3 F(1, \tau; E_3) \leq \lambda S_1 F(1, \tau; E_1) + (1 - \lambda)S_2 F(1, \tau; E_2).$$

From Theorem 9, F is homogeneous of degree one in S and E . Hence,

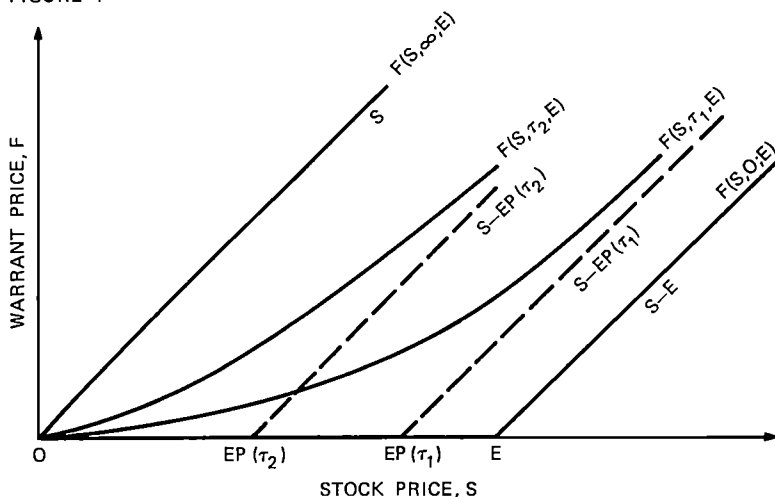
$$F(S_3, \tau; S_3 E_3) \leq \lambda F(S_1, \tau; S_1 E_1) + (1 - \lambda)F(S_2, \tau; S_2 E_2).$$

By the definition of E_1 , E_2 , and E_3 , this inequality can be rewritten as $F(S_3, \tau; E) \leq \lambda F(S_1, \tau; E) + (1 - \lambda)F(S_2, \tau; E)$. Q.E.D.

Although convexity is usually assumed to be a property which always holds for warrants, and while the hypothesized conditions of Theorem 10 are by no means necessary, Appendix 1 provides an example where the distribution of future returns on the common stock is sufficiently dependent on the level of the stock price, to cause perverse local concavity.

Based on the analysis so far, Figure 1 illustrates the general shape that the rational warrant price should satisfy as a function of the stock price and time.

FIGURE 1



■ A number of the theorems of the previous section depend upon the assumption that either no payouts are made to the common stock over the life of the contract or that the contract is protected against such payments. In this section, the adjustments required in the contracts to protect them against payouts are derived, and the effects of payouts on the valuation of unprotected contracts are investigated. The two most common types of payouts are stock dividends (splits) and cash dividends.

In general, the value of an option will be affected by unanticipated changes in the firm's investment policy, capital structure (e.g., debt-equity ratio), and payout policy. For example, if the firm should change its investment policy so as to lower the riskiness of its cash flow (and hence, the riskiness of outcomes on the common stock), then, by Theorem 8, the value of the warrant would decline for a given level of the stock price. Similarly, if the firm changed its capital structure by raising the debt-equity ratio, then the riskiness of the common stock would increase, and the warrant would become more valuable. If that part of the total return received by shareholders in the form of dividends is increased by a change in payout policy, then the value of an unprotected warrant would decline since the warrant-holder has no claim on the dividends.¹⁸

While it is difficult to provide a set of adjustments to the warrant contract to protect it against changes in investment or capital structure policies without severely restricting the management of the firm, there do exist a set of adjustments to protect the warrant holders against payouts.

Definition: An option is said to be *payout protected* if, for a fixed investment policy and fixed capital structure, the value of the option is invariant to the choice of payout policy.

Theorem 11. If the total return per dollar invested in the common stock is invariant to the fraction of the return represented by payouts and if, on each expayout date during the life of a warrant, the contract is adjusted so that the number of shares which can be purchased for a total of E dollars is increased by (d/S^x) percent where d is the dollar amount of the payout and S^x is the expayout price per share of the stock, then the warrant will be payout protected.

Proof: Consider two firms with identically distributed total returns per dollar invested in the common stock, $z_i(t)$, $i = 1, 2$, and whose initial prices per share are the same ($S_1 = S_2 = S$). For firm i , let $\lambda_i(t)$ ($t \geq 1$) be the return per dollar in period t from payouts and $x_i(t)$ be the return per dollar in period t from capital gains, such that $z_i(t) \equiv \lambda_i(t)x_i(t)$. Let $N_i(t)$ be the number of shares of firm i which the warrant of firm i has claim on for a total price of E , at time t where $N_1(0) = N_2(0) = 1$. By definition, $\lambda_i(t) \equiv 1 + d_i(t)/S_i^x(t)$, where $S_i^x(t) = \prod_{k=1}^t x_i(k)S$ is the expayout price per share at time t . Therefore, by the hypothesized conditions of the theorem, $N_i(t) = \lambda_i(t)N_i(t-1)$. On the date when the warrants are either exercised

¹⁸ This is an important point to remember when valuing unprotected warrants of companies such as A. T. & T. where a substantial fraction of the total return to shareholders comes in the form of dividends.

or expire, the value of the warrant on firm i will be

$$\text{Max}[0, N_i(t)S_i^x(t) - E].$$

But, $N_i(t)S_i^x(t) = [\prod_{k=1}^t \lambda_i(t)] [\prod_{k=1}^t x_i(t)S] = \prod_{k=1}^t z_i(t)S$. Since, by hypothesis, the $z_i(t)$ are identically distributed, the distribution of outcomes on the warrants of the two firms will be identical. Therefore, by Assumption 2, $F_1(S, \tau; E) = F_2(S, \tau; E)$, independent of the particular pattern chosen for the $\lambda_i(t)$. Q.E.D.

Note that if the hypothesized conditions of Theorem 11 hold, then the value of a protected warrant will be equal to the value of a warrant which restricts management from making any payouts to the common stock over the life of the warrant (i.e., $\lambda_i(t) \equiv 1$). Hence, a protected warrant will satisfy all the theorems of Section 2 which depend on the assumption of no payouts over the life of the warrant.

Corollary. If the total return per dollar invested in the common stock is invariant to the fraction of the return represented by payouts; if there are no economies of scale; and if, on each expayout date during the life of a warrant, each warrant to purchase one share of stock for exercise price E , is exchanged for λ ($\equiv 1 + d/S^x$) warrants to purchase one share of stock for exercise price E/λ , then the warrant will be payout protected.

Proof: By Theorem 11, on the first expayout date, a protected warrant will have claim on λ shares of stock at a total exercise price of E . By hypothesis, there are no economies of scale. Hence, the scale interpretation of Theorem 6 is valid which implies that the value of a warrant on λ shares at a total price of E must be identically (in λ) equal to the value of λ warrants to purchase one share at an exercise price of E/λ . Proceeding inductively, we can show that this equality holds on each payout date. Hence, a warrant with the adjustment provision of the Corollary will be payout protected. Q.E.D.

If there are no economies of scale, it is generally agreed that a stock split or dividend will not affect the distribution of future per dollar returns on the common stock. Hence, the hypothesized adjustments will protect the warrant holder against stock splits where λ is the number of postsplit shares per presplit share.¹⁹

The case for cash dividend protection is more subtle. In the absence of taxes and transactions costs, Miller and Modigliani²⁰ have shown that for a fixed investment policy and capital structure, dividend policy does not affect the value of the firm. Under their hypothesized conditions, it is a necessary result of their analysis that the total return per dollar invested in the common stock will be invariant to payout policy. Therefore, warrants adjusted according to either Theorem 11 or its Corollary, will be payout protected in the same

¹⁹ For any particular function, $F(S, \tau; E)$, there are many other adjustments which could leave value the same. However, the adjustment suggestions of Theorem 11 and its Corollary are the only ones which do so for every such function. In practice, both adjustments are used to protect warrants against stock splits. See Braniff Airways 1986 warrants for an example of the former and Leasco 1987 warrants for the latter. λ could be less than one in the case of a reverse split.

²⁰ In [35].

sense that Miller and Modigliani mean when they say that dividend policy “doesn’t matter.”

The principal cause for confusion is different definitions of payout protected. Black and Scholes²¹ give an example to illustrate “that there may not be any adjustment in the terms of the option that will give adequate protection against a large dividend.” Suppose that the firm liquidates all its assets and pays them out in the form of a cash dividend. Clearly, $S^x = 0$, and hence, the value of the warrant must be zero no matter what adjustment is made to the number of shares it has claim on or to its exercise price.

While their argument is correct, it also suggests a much stronger definition of payout protection. Namely, since their example involves changes in investment policy and if there is a positive supply of warrants (the nonincipient case), a change in the capital structure, in addition to a payout, their definition would seem to require protection against all three.

To illustrate, consider the firm in their example, but where management is prohibited against making any payouts to the shareholders prior to expiration of the warrant. It seems that such a warrant would be called payout protected by any reasonable definition. It is further assumed that the firm has only equity outstanding (i.e., the incipient case for the warrant) to rule out any capital structure effects.²²

Suppose the firm sells all its assets for a fair price (so that the share price remains unchanged) and uses the proceeds to buy riskless, τ -period bonds. As a result of this investment policy change, the stock becomes a riskless asset and the warrant price will fall to $\text{Max}[0, S - EP]$. Note that if $S < EP$, the warrant will be worthless even though it is payout protected. Now lift the restriction against payouts and replace it with the adjustments of the Corollary to Theorem 11. Given that the shift in investment policy has taken place, suppose the firm makes a payment of γ percent of the value of the firm to the shareholders. Then, $S^x = (1 - \gamma)S$ and

$$\lambda = 1 + \gamma/(1 - \gamma) = 1/(1 - \gamma).$$

The value of the warrant after the payout will be

$$\lambda \text{Max}[0, S^x - EP/\lambda] = \text{Max}[0, S - EP],$$

which is the same as the value of the warrant when the company was restricted from making payouts. In the B-S example, $\gamma = 1$ and so, $\lambda = \infty$ and $E/\lambda = 0$. Hence, there is the indeterminacy of multiplying zero by infinity. However, for every $\gamma < 1$, the analysis is correct, and therefore, it is reasonable to suspect that it holds in the limit.

A similar analysis in the nonincipient case would show that both investment policy and the capital structure were changed. For in this case, the firm would have to purchase γ percent of the warrants outstanding to keep the capital structure unchanged without issuing new stock. In the B-S example where $\gamma = 1$, this would require purchasing

²¹ In [4].

²² The incipient case is a particularly important example since in practice, the only contracts that are adjusted for cash payouts are options. The incipient assumption also rules out “capital structure induced” changes in investment policy by malevolent management. For an example, see Stiglitz [50].

the entire issue, after which the analysis reduces to the incipient case. The B-S emphasis on protection against a "large" dividend is further evidence that they really have in mind protection against investment policy and capital structure shifts as well, since large payouts are more likely to be associated with nontrivial changes in either or both.

It should be noted that calls and puts that satisfy the incipient assumption have in practice been the only options issued with cash dividend protection clauses, and the typical adjustment has been to reduce the exercise price by the amount of the cash dividend which has been demonstrated to be incorrect.²³

To this point it has been assumed that the exercise price remains constant over the life of the contract (except for the before-mentioned adjustments for payouts). A variable exercise price is meaningless for an European warrant since the contract is not exercisable prior to expiration. However, a number of American warrants do have variable exercise prices as a function of the length of time until expiration. Typically, the exercise price increases as time approaches the expiration date.

Consider the case where there are n changes of the exercise price during the life of an American warrant, represented by the following schedule:

<i>Exercise Price</i>	<i>Time until Expiration (τ)</i>
E_0	$0 \leq \tau \leq \tau_1$
E_1	$\tau_1 \leq \tau \leq \tau_2$
$\vdots$	$\vdots$
E_n	$\tau_n \leq \tau,$

where it is assumed that $E_{j+1} < E_j$ for $j = 0, 1, \dots, n-1$. If, otherwise the conditions for Theorems 1-11 hold, it is easy to show that, if premature exercising takes place, it will occur only at points in time just prior to an exercise price change, i.e., at $\tau = \tau_j^+$, $j = 1, 2, \dots, n$. Hence, the American warrant is equivalent to a *modified European warrant* which allows its owner to exercise the warrant at discrete times, just prior to an exercise price change. Given a technique for finding the price of an European warrant, there is a systematic method for valuing a modified European warrant. Namely, solve the standard problem for $F_0(S, \tau; E_0)$ subject to the boundary conditions $F_0(S, 0; E_0) = \text{Max}[0, S - E_0]$ and $\tau \leq \tau_1$. Then, by the same technique, solve for $F_1(S, \tau; E_1)$ subject to the boundary conditions $F_1(S, \tau_1; E_1) = \text{Max}[0, S - E_1, F_0(S, \tau_1; E_0)]$ and $\tau_1 \leq \tau \leq \tau_2$. Proceed inductively by this dynamic-programming-like technique, until the current value of the modified European warrant is determined. Typically, the number of exercise price changes is small, so the technique is computationally feasible.

Often the contract conditions are such that the warrant will never be prematurely exercised, in which case, the correct valuation will be the standard European warrant treatment using the exercise

²³ By Taylor series approximation, we can compute the loss to the warrant holder of the standard adjustment for dividends: namely, $F(S - d, \tau; E - d) - F(S, \tau; E) = -dF_S(S, \tau; E) - dF_E(S, \tau; E) + o(d) = -[F(S, \tau; E) - (S - E)F_S(S, \tau; E)](d/E) + o(d)$, by the first-degree homogeneity of F in (S, E) . Hence, to a first approximation, for $S = E$, the warrant will lose (d/S) percent of its value by this adjustment. Clearly, for $S > E$, the percentage loss will be smaller and for $S < E$, it will be larger.

price at expiration, E_0 . If it can be demonstrated that

$$F_j(S, \tau_{j+1}; E_j) \geq S - E_{j+1} \quad \text{for all } S \geq 0 \text{ and } j = 0, 1, \dots, N-1, \quad (10)$$

then the warrant will always be worth more "alive" than "dead," and the no-premature exercising result will obtain. From Theorem 1, $F_j(S, \tau_{j+1}; E_j) \geq \text{Max}[0, S - P(\tau_{j+1} - \tau_j)E_j]$. Hence, from (10), a sufficient condition for no early exercising is that

$$E_{j+1}/E_j > P(\tau_{j+1} - \tau_j). \quad (11)$$

The economic reasoning behind (11) is identical to that used to derive Theorem 1. If by continuing to hold the warrant and investing the dollars which would have been paid for the stock if the warrant were exercised, the investor can with certainty earn enough to overcome the increased cost of exercising the warrant later, then the warrant should not be exercised.

Condition (11) is not as simple as it may first appear, because in valuing the warrant today, one must know for certain that (11) will be satisfied at some future date, which in general will not be possible if interest rates are stochastic. Often, as a practical matter, the size of the exercise price change versus the length of time between changes is such that for almost any reasonable rate of interest, (11) will be satisfied. For example, if the increase in exercise price is 10 percent and the length of time before the next exercise price change is five years, the yield to maturity on riskless securities would have to be less than 2 percent before (11) would not hold.

As a footnote to the analysis, we have the following Corollary.

Corollary. If there is a finite number of changes in the exercise price of a payout-protected, perpetual warrant, then it will not be exercised and its price will equal the common stock price.

Proof: applying the previous analysis, consider the value of the warrant if it survives past the last exercise price change, $F_0(S, \infty; E_0)$. By Theorem 3, $F_0(S, \infty; E_0) = S$. Now consider the value just prior to the last change in exercise price, $F_1(S, \infty; E_1)$. It must satisfy the boundary condition,

$$\begin{aligned} F_1(S, \infty; E_1) &= \text{Max}[0, S - E_1, F_0(S, \infty; E_0)] \\ &= \text{Max}[0, S - E_1, S] = S. \end{aligned}$$

Proceeding inductively, the warrant will never be exercised, and by Theorem 3, its value is equal to the common stock. Q.E.D.

The analysis of the effect on unprotected warrants when future dividends or dividend policy is known,²⁴ follows exactly the analysis of a changing exercise price. The arguments that no one will prematurely exercise his warrant except possibly at the discrete points in time just prior to a dividend payment, go through, and hence, the modified European warrant approach works where now the boundary conditions are $F_j(S, \tau_j; E) = \text{Max}[0, S - E, F_{j-1}(S - d_j, \tau_j; E)]$

²⁴ The distinction is made between knowing future dividends and dividend policy. With the former, one knows, currently, the actual amounts of future payments while, with the latter, one knows the conditional future payments, conditional on (currently unknown) future values, such as the stock price.

where d , equals the dividend per share paid at τ , years prior to expiration, for $j = 1, 2, \dots, n$.

In the special case, where future dividends and rates of interest are known with certainty, a sufficient condition for no premature exercising is that²⁵

$$E > \sum_{t=0}^{\tau} d(t)P(\tau - t)/[1 - P(\tau)]. \quad (12)$$

I.e., the net present value of future dividends is less than the present value of earnings from investing E dollars for τ periods. If dividends are paid continuously at the constant rate of d dollars per unit time and if the interest rate, r , is the same over time, then (12) can be rewritten in its continuous form as

$$E > \frac{d}{r}. \quad (13)$$

Samuelson suggests the use of discrete recursive relationships, similar to our modified European warrant analysis, as an approximation to the mathematically difficult continuous-time model when there is some chance for premature exercising.²⁶ We have shown that the only reasons for premature exercising are lack of protection against dividends or sufficiently unfavorable exercise price changes. Further, such exercising will never take place except at boundary points. Since dividends are paid quarterly and exercise price changes are less frequent, the Samuelson recursive formulation with the discrete-time spacing matching the intervals between dividends or exercise price changes is actually the correct one, and the continuous solution is the approximation, even if warrant and stock prices change continuously!

Based on the relatively weak Assumption 1, we have shown that dividends and unfavorable exercise price changes are the only rational reasons for premature exercising, and hence, the only reasons for an American warrant to sell for a premium over its European counterpart. In those cases where early exercising is possible, a computationally feasible, general algorithm for modifying a European warrant valuation scheme has been derived. A number of theorems were proved putting restrictions on the structure of rational European warrant pricing theory.

4. Restrictions on rational put option pricing

■ The put option, defined at the beginning of Section 2, has received relatively little analysis in the literature because it is a less popular option than the call and because it is commonly believed²⁷ that, given the price of a call option and the common stock, the value of a put is uniquely determined. This belief is false for American put

²⁵ The interpretation of (12) is similar to the explanation given for (11). Namely, if the losses from dividends are smaller than the gains which can be earned risklessly, from investing the extra funds required to exercise the warrant and hold the stock, then the warrant is worth more "alive" than "dead."

²⁶ See [42], pp. 25–26, especially equation (42). Samuelson had in mind small, discrete-time intervals, while in the context of the current application, the intervals would be large. Chen [8] also used this recursive relationship in his empirical testing of the Samuelson model.

²⁷ See, for example, Black and Scholes [4] and Stoll [52].

options, and the mathematics of put options pricing is more difficult than that of the corresponding call option.

Using the notation defined in Section 2, we have that, at expiration,

$$G(S, 0; E) = g(S, 0; E) = \text{Max} [0, E - S]. \quad (14)$$

To determine the rational European put option price, two portfolio positions are examined. Consider taking a long position in the common stock at S dollars, a long position in a τ -year European put at $g(S, \tau; E)$ dollars, and borrowing $[EP'(\tau)]$ dollars where $P'(\tau)$ is the current value of a dollar payable τ -years from now at the borrowing rate²⁸ (i.e., $P'(\tau)$ may not equal $P(\tau)$ if the borrowing and lending rates differ). The value of the portfolio τ years from now with the stock price at S^* will be: $S^* + (E - S^*) - E = 0$, if $S^* \leq E$, and $S^* + 0 - E = S^* - E$, if $S^* > E$. The pay-off structure is identical in every state to a European call option with the same exercise price and duration. Hence, to avoid the call option from being a dominated security,²⁹ the put and call must be priced so that

$$g(S, \tau; E) + S - EP'(\tau) \geq f(S, \tau; E). \quad (15)$$

As was the case in the similar analysis leading to Theorem 1, the values of the portfolio prior to expiration were not computed because the call option is European and cannot be prematurely exercised.

Consider taking a long position in a τ -year European call, a short position in the common stock at price S , and lending $EP(\tau)$ dollars. The value of the portfolio τ years from now with the stock price at S^* will be: $0 - S^* + E = E - S^*$, if $S^* \leq E$, and $(S^* - E) - S^* + E = 0$, if $S^* > E$. The pay-off structure is identical in every state to a European put option with the same exercise price and duration. If the put is not to be a dominated security,³⁰ then

$$f(S, \tau; E) - S + EP(\tau) \geq g(S, \tau; E) \quad (16)$$

must hold.

Theorem 12. If Assumption 1 holds and if the borrowing and lending rates are equal [i.e., $P(\tau) = P'(\tau)$], then

$$g(S, \tau; E) = f(S, \tau; E) - S + EP(\tau).$$

Proof: the proof follows directly from the simultaneous application of (15) and (16) when $P'(\tau) = P(\tau)$. Q.E.D.

Thus, the value of a rationally priced European put option is determined once one has a rational theory of the call option value. The formula derived in Theorem 12 is identical to B-S's equation (26), when the riskless rate, r , is constant (i.e., $P(\tau) = e^{-r\tau}$). Note

²⁸ The borrowing rate is the rate on a τ -year, noncallable, discounted loan. To avoid arbitrage, $P'(\tau) \leq P(\tau)$.

²⁹ Due to the existent market structure, (15) must hold for the stronger reason of arbitrage. The portfolio did not require short-sales and it is institutionally possible for an investor to issue (sell) call options and reinvest the proceeds from the sale. If (15) did not hold, an investor, acting unilaterally, could make immediate, positive profits with no investment and no risk.

³⁰ In this case, we do not have the stronger condition of arbitrage discussed in footnote (29) because the portfolio requires a short sale of shares, and, under current regulations, the proceeds cannot be reinvested. Again, intermediate values of the portfolio are not examined because the put option is European.

that no distributional assumptions about the stock price or future interest rates were required to prove Theorem 12.

Two corollaries to Theorem 12 follow directly from the above analysis.

Corollary 1. $EP(\tau) \geq g(S, \tau; E)$.

Proof: from (5) and (7), $f(S, \tau; E) - S \leq 0$ and from (16), $EP(\tau) \geq g(S, \tau; E)$. Q.E.D.

The intuition of this result is immediate. Because of limited liability on the common stock, the maximum value of the put option is E , and because the option is European, the proceeds cannot be collected for τ years. The option cannot be worth more than the present value of a sure payment of its maximum value.

Corollary 2. The value of a perpetual ($\tau = \infty$) European put option is zero.

Proof: the put is a limited liability security [$g(S, \tau; E) \geq 0$]. From Corollary 1 and the condition that $P(\infty) = 0$, $0 \geq g(S, \infty; E)$. Q.E.D.

Using the relationship $g(S, \tau; E) = f(S, \tau; E) - S + EP(\tau)$, it is straightforward to derive theorems for rational European put pricing which are analogous to the theorems for warrants in Section 2. In particular, whenever f is homogeneous of degree one or convex in S and E , so g will be also. The correct adjustment for stock and cash dividends is the same as prescribed for warrants in Theorem 11 and its Corollary.³¹

Since the American put option can be exercised at any time, its price must satisfy the arbitrage condition

$$G(S, \tau; E) \geq \text{Max}[0, E - S]. \quad (17)$$

By the same argument used to derive (5), it can be shown that

$$G(S, \tau; E) \geq g(S, \tau; E), \quad (18)$$

where the strict inequality holds only if there is a positive probability of premature exercising.

As shown in Section 2, the European and American warrant have the same value if the exercise price is constant and they are protected against payouts to the common stock. Even under these assumptions, there is almost always a positive probability of premature exercising of an American put, and hence, the American put will sell for more than its European counterpart. A hint that this must be so comes from Corollary 2 and arbitrage condition (17). Unlike European options, the value of an American option is always a nondecreasing function of its expiration date. If there is no possibility of premature exercising, the value of an American option will equal the value of its European counterpart. By the Corollary to Theorem 11, the value of a perpetual American put would be zero, and by the monotonicity argument on length of time to maturity, all American puts would have zero value.

³¹ While such adjustments for stock or cash payouts add to the value of a warrant or call option, the put option owner would prefer not to have them since lowering the exercise price on a put decreases its value. For simplicity, the effects of payouts are not considered, and it is assumed that no dividends are paid on the stock, and there are no exercise price changes.

This absurd result clearly violates the arbitrage condition (17) for $S < E$.

To clarify this point, reconsider the two portfolios examined in the European put analysis, but with American puts instead. The first portfolio contained a long position in the common stock at price S , a long position in an American put at price $G(S, \tau; E)$, and borrowings of $[EP'(\tau)]$. As was previously shown, if held until maturity, the outcome of the portfolio will be identical to those of an American (European) warrant held until maturity. Because we are now using American options with the right to exercise prior to expiration, the interim values of the portfolio must be examined as well. If, for all times prior to expiration, the portfolio has value greater than the exercise value of the American warrant, $S - E$, then to avoid dominance of the warrant, the current value of the portfolio must exceed or equal the current value of the warrant.

The interim value of the portfolio at T years until expiration when the stock price is S^* , is

$$\begin{aligned} S^* + G(S^*, T; E) - EP'(T) \\ = G(S^*, T; E) + (S^* - E) + E[1 - P'(T)] > (S^* - E). \end{aligned}$$

Hence, condition (15) holds for its American counterparts to avoid dominance of the warrant, i.e.,

$$G(S, \tau; E) + S - EP'(\tau) \geq F(S, \tau; E). \quad (19)$$

The second portfolio has a long position in an American call at price $F(S, \tau; E)$, a short position in the common stock at price S , and a loan of $[EP(\tau)]$ dollars. If held until maturity, this portfolio replicates the outcome of a European put, and hence, must be at least as valuable at any interim point in time. The interim value of the portfolio, at T years to go and with the stock price at S^* , is

$$\begin{aligned} F(S^*, T; E) - S^* + EP(T) \\ = (E - S^*) + F(S^*, T; E) - E[1 - P(T)] < E - S^*, \end{aligned}$$

if $F(S^*, T; E) < E[1 - P(T)]$, which is possible for small enough S^* . From (17), $G(S^*, T; E) \geq E - S^*$. So, the interim value of the portfolio will be less than the value of an American put for sufficiently small S^* . Hence, if an American put was sold against this portfolio, and if the put owner decided to exercise his put prematurely, the value of the portfolio could be less than the value of the exercised put. This result would certainly obtain if $S^* < E[1 - P(T)]$. So, the portfolio will not dominate the put if inequality (16) does not hold, and an analog theorem to Theorem 12, which uniquely determines the value of an American put in terms of a call, does not exist. Analysis of the second portfolio does lead to the weaker inequality that

$$G(S, \tau; E) \leq E - S + F(S, \tau; E). \quad (20)$$

Theorem 13. If, for some $T < \tau$, there is a positive probability that $f(S, T; E) < E[1 - P(T)]$, then there is a positive probability that a τ -year, American put option will be exercised prematurely and the value of the American put will strictly exceed the value of its European counterpart.

Proof: the only reason that an American put will sell for a premium over its European counterpart is that there is a positive probability

of exercising prior to expiration. Hence, it is sufficient to prove that $g(S, \tau; E) < G(S, \tau; E)$. From Assumption 1, if for some $T \leq \tau$, $g(S^*, T; E) < G(S^*, T; E)$ for some possible value(s) of S^* , then $g(S, \tau; E) < G(S, \tau; E)$. From Theorem 12, $g(S^*, T; E) = f(S^*, T; E) - S^* + EP(T)$. From (17), $G(S^*, T; E) \geq \text{Max}[0, E - S^*]$. But $g(S^*, T; E) < G(S^*, T; E)$ is implied if $E - S^* > f(S^*, T; E) - S^* + EP(T)$, which holds if $f(S^*, T; E) < E[1 - P(T)]$. By hypothesis of the theorem, such an S^* is a possible value. Q.E.D.

Since almost always there will be a chance of premature exercising, the formula of Theorem 12 or B-S equation (26) will not lead to a correct valuation of an American put and, as mentioned in Section 3, the valuation of such options is a more difficult analytical task than valuing their European counterparts.

5. Rational option pricing along Black-Scholes lines

■ A number of option pricing theories satisfy the general restrictions on a rational theory as derived in the previous sections. One such theory developed by B-S³² is particularly attractive because it is a complete general equilibrium formulation of the problem and because the final formula is a function of “observable” variables, making the model subject to direct empirical tests.

B-S assume that: (1) the standard form of the Sharpe-Lintner-Mossin capital asset pricing model holds for intertemporal trading, and that trading takes place continuously in time; (2) the market rate of interest, r , is known and fixed over time; and (3) there are no dividends or exercise price changes over the life of the contract.

To derive the formula, they assume that the option price is a function of the stock price and time to expiration, and note that, over “short” time intervals, the stochastic part of the change in the option price will be perfectly correlated with changes in the stock price. A hedged portfolio containing the common stock, the option, and a short-term, riskless security, is constructed where the portfolio weights are chosen to eliminate all “market risk.” By the assumption of the capital asset pricing model, any portfolio with a zero (“beta”) market risk must have an expected return equal to the risk-free rate. Hence, an equilibrium condition is established between the expected return on the option, the expected return on the stock, and the riskless rate.

Because of the distributional assumptions and because the option price is a function of the common stock price, B-S in effect make use of the Samuelson³³ application to warrant pricing of the Bachelier-Einstein-Dynkin derivation of the Fokker-Planck equation, to express the expected return on the option in terms of the option price function and its partial derivatives. From the equilibrium condition on the option yield, such a partial differential equation for the option price is derived. The solution to this equation for a European call option is

$$f(S, \tau; E) = S\Phi(d_1) - Ee^{-r\tau}\Phi(d_2), \quad (21)$$

where Φ is the cumulative normal distribution function, σ^2 is the

³² In [4].

³³ In [42].

instantaneous variance of the return on the common stock,

$$d_1 \equiv [\log(S/E) + (r + \frac{1}{2}\sigma^2)\tau]/\sigma\sqrt{\tau},$$

and $d_2 \equiv d_1 - \sigma\sqrt{\tau}$.

An exact formula for an asset price, based on observable variables only, is a rare finding from a general equilibrium model, and care should be taken to analyze the assumptions with Occam's razor to determine which ones are necessary to derive the formula. Some hints are to be found by inspection of their final formula (21) and a comparison with an alternative general equilibrium development.

The manifest characteristic of (21) is the number of variables that it does *not* depend on. The option price does not depend on the expected return on the common stock,³⁴ risk preferences of investors, or on the aggregate supplies of assets. It does depend on the rate of interest (an "observable") and the *total* variance of the return on the common stock which is often a stable number and hence, accurate estimates are possible from time series data.

The Samuelson and Merton³⁵ model is a complete, although very simple (three assets and one investor) general equilibrium formulation. Their formula³⁶ is

$$f(S, \tau; E) = e^{-r\tau} \int_{E/S}^{\infty} (ZS - E)dQ(Z; \tau), \quad (22)$$

where dQ is a probability density function with the expected value of Z over the dQ distribution equal to $e^{r\tau}$. Equations (22) and (21) will be the same only in the special case when dQ is a log-normal density with the variance of $\log(Z)$ equal to $\sigma^2\tau$.³⁷ However, dQ is a risk-adjusted ("util-prob") distribution, dependent on both risk-preferences and aggregate supplies, while the distribution in (21) is the objective distribution of returns on the common stock. B-S claim that one reason that Samuelson and Merton did not arrive at formula (21) was because they did not consider other assets. If a result does not obtain for a simple, three asset case, it is unlikely that it would in a more general example. More to the point, it is only necessary to consider three assets to derive the B-S formula. In connection with this point, although B-S claim that their central assumption is the capital asset pricing model (emphasizing this over their hedging argument), their final formula, (21), depends only on the interest rate (which is exogenous to the capital asset pricing model) and on the *total* variance of the return on the common stock. It does not depend on the betas (covariances with the market) or other assets' characteristics. Hence, this assumption may be a "red herring."

Although their derivation of (21) is intuitively appealing, such an

³⁴ This is an important result because the expected return is not directly observable and estimates from past data are poor because of nonstationarity. It also implies that attempts to use the option price to estimate expected returns on the stock or risk-preferences of investors are doomed to failure (e.g., see Sprenkle [49]).

³⁵ In [43].

³⁶ *Ibid.*, p. 29, equation 30.

³⁷ This will occur only if: (1) the objective returns on the stock are log-normally distributed; (2) the investor's utility function is iso-elastic (i.e., homothetic indifference curves); and (3) the supplies of *both* options and bonds are at the incipient level.

important result deserves a rigorous derivation. In this case, the rigorous derivation is not only for the satisfaction of the “purist,” but also to give insight into the necessary conditions for the formula to obtain. The reader should be alerted that because B-S consider only terminal boundary conditions, their analysis is strictly applicable to European options, although as shown in Sections 2 through 4, the European valuation is often equal to the American one.

Finally, although their model is based on a different economic structure, the formal analytical content is identical to Samuelson’s “linear, $\alpha = \beta$ ” model when the returns on the common stock are log-normal.³⁸ Hence, with different interpretation of the parameters, theorems proved in Samuelson and in the difficult McKean appendix³⁹ are directly applicable to the B-S model, and vice versa.

6. An alternative derivation of the Black-Scholes model⁴⁰

■ Initially, we consider the case of a European option where no payouts are made to the common stock over the life of the contract. We make the following further assumptions.

- (1) “Frictionless” markets: there are no transactions costs or differential taxes. Trading takes place continuously and borrowing and short-selling are allowed without restriction.⁴¹ The borrowing rate equals the lending rate.
- (2) *Stock price dynamics*: the instantaneous return on the common stock is described by the stochastic differential equation⁴²

$$\frac{dS}{S} = \alpha dt + \sigma dz, \quad (23)$$

where α is the instantaneous expected return on the common stock, σ^2 is the instantaneous variance of the return, and dz is a standard Gauss-Wiener process. α may be a stochastic variable of quite general type including being dependent on the level of the stock price or other assets’ returns. Therefore, no presumption is made that dS/S is an independent increments process or stationary, although dz clearly is. However,

³⁸ In [42]. See Merton [28] for a brief description of the relationship between the Samuelson and B-S models.

³⁹ In [26].

⁴⁰ Although the derivation presented here is based on assumptions and techniques different from the original B-S model, it is in the spirit of their formulation, and yields the same formula when their assumptions are applied.

⁴¹ The assumptions of unrestricted borrowing and short-selling can be weakened and still have the results obtained by splitting the created portfolio of the text into two portfolios: one containing the common stock and the other containing the warrant plus a long position in bonds. Then, as was done in Section 2, if we accept Assumption 1, the formulas of the current section follow immediately.

⁴² For a general description of the theory of stochastic differential equations of the Itô type, see McKean [27] and Kushner [24]. For a description of their application to the consumption-portfolio problem, see Merton [32], [33], and [31]. Briefly, Itô processes follow immediately from the assumption of a continuous-time stochastic process which results in continuous price changes (with finite moments) and some level of independent increments. If the process for price changes were functions of stable Paretian distributions with infinite moments, it is conjectured that the only equilibrium value for a warrant would be the stock price itself, independent of the length of time to maturity. This implication is grossly inconsistent with all empirical observations.

σ is restricted to be nonstochastic and, at most, a known function of time.

- (3) *Bond price dynamics:* $P(\tau)$ is as defined in previous sections and the dynamics of its returns are described by

$$\frac{dP}{P} = \mu(\tau)dt + \delta(\tau)dq(t; \tau), \quad (24)$$

where μ is the instantaneous expected return, δ^2 is the instantaneous variance, and $dq(t; \tau)$ is a standard Gauss-Wiener process for maturity τ . Allowing for the possibility of habitat and other term structure effects, it is not assumed that dq for one maturity is perfectly correlated with dq for another, i.e.,

$$dq(t; \tau)dq(t; T) = \rho_{\tau T}dt, \quad (24a)$$

where $\rho_{\tau T}$ may be less than one for $\tau \neq T$. However, it is assumed that there is no serial correlation⁴³ among the (unanticipated) returns on any of the assets, i.e.,

$$\begin{aligned} dq(s; \tau)dq(t; T) &= 0 \quad \text{for } s \neq t \\ dq(s; \tau)dz(t) &= 0 \quad \text{for } s \neq t, \end{aligned} \quad (24b)$$

which is consistent with the general efficient market hypothesis of Fama and Samuelson.⁴⁴ $\mu(\tau)$ may be stochastic through dependence on the level of bond prices, etc., and different for different maturities. Because $P(\tau)$ is the price of a discounted loan with no risk of default, $P(0) = 1$ with certainty and $\delta(\tau)$ will definitely depend on τ with $\delta(0) = 0$. However, δ is otherwise assumed to be nonstochastic and independent of the level of P . In the special case when the interest rate is nonstochastic and constant over time, $\delta \equiv 0$, $\mu = r$, and $P(\tau) = e^{-r\tau}$.

- (4) *Investor preferences and expectations:* no assumptions are necessary about investor preferences other than that they

⁴³ The reader should be careful to note that it is assumed only that the *unanticipated* returns on the bonds are not serially correlated. Cootner [11] and others have pointed out that since the bond price will equal its redemption price at maturity, the total returns over time cannot be uncorrelated. In no way does this negate the specification of (24), although it does imply that the variance of the unanticipated returns must be a function of time to maturity. An example to illustrate that the two are not inconsistent can be found in Merton [29]. Suppose that bond prices for all maturities are only a function of the current (and future) short-term interest rates. Further, assume that the short-rate, r , follows a Gauss-Wiener process with (possibly) some drift, i.e., $dr = adt + gdz$, where a and g are constants. Although this process is not realistic because it implies a positive probability of negative interest rates, it will still illustrate the point. Suppose that all bonds are priced so as to yield an expected rate of return over the next period equal to r (i.e., a form of the expectations hypothesis):

$$P(\tau; r) = \exp \left[-r\tau - \frac{a}{2} \tau^2 + \frac{g^2 \tau^3}{6} \right]$$

and

$$\frac{dP}{P} = rdt - g\tau dz.$$

By construction, dz is not serially correlated and in the notation of (24), $\delta(\tau) = -g\tau$.

⁴⁴ In [13] and [41], respectively.

satisfy Assumption 1 of Section 2. All investors agree on the values of σ and δ , and on the distributional characteristics of dz and dq . It is *not* assumed that they agree on either α or μ .⁴⁵

From the analysis in Section 2, it is reasonable to assume that the option price is a function of the stock price, the riskless bond price, and the length of time to expiration. If $H(S, P, \tau; E)$ is the option price function, then, given the distributional assumptions on S and P , we have, by Itô's Lemma,⁴⁶ that the change in the option price over time satisfies the stochastic differential equation,

$$dH = H_1 dS + H_2 dP + H_3 d\tau + \frac{1}{2}[H_{11}(dS)^2 + 2H_{12}(dSdP) + H_{22}(dP)^2], \quad (25)$$

where subscripts denote partial derivatives, and $(dS)^2 \equiv \sigma^2 S^2 dt$, $(dP)^2 \equiv \delta^2 P^2 dt$, $d\tau = -dt$, and $(dSdP) \equiv \rho\sigma\delta SPdt$ with ρ , the instantaneous correlation coefficient between the (unanticipated) returns on the stock and on the bond. Substituting from (23) and (24) and rearranging terms, we can rewrite (25) as

$$dH = \beta H dt + \gamma H dz + \eta H dq, \quad (26)$$

where the instantaneous expected return on the warrant, β , equals $[\frac{1}{2}\sigma^2 S^2 H_{11} + \rho\sigma\delta SP H_{12} + \frac{1}{2}\delta^2 P^2 H_{22} + \alpha SH_1 + \mu PH_2 - H_3]/H$, $\gamma \equiv \sigma SH_1/H$, and $\eta \equiv \delta PH_2/H$.

In the spirit of the Black-Scholes formulation and the analysis in Sections 2 thru 4, consider forming a portfolio containing the common stock, the option, and riskless bonds with time to maturity, τ , equal to the expiration date of the option, such that the aggregate investment in the portfolio is zero. This is achieved by using the proceeds of short-sales and borrowing to finance long positions. Let W_1 be the (instantaneous) number of dollars of the portfolio invested in the common stock, W_2 be the number of dollars invested in the option, and W_3 be the number of dollars invested in bonds. Then, the condition of zero aggregate investment can be written as $W_1 + W_2 + W_3 = 0$. If dY is the instantaneous dollar return to the portfolio, it can be shown⁴⁷ that

$$\begin{aligned} dY &= W_1 \frac{dS}{S} + W_2 \frac{dH}{H} + W_3 \frac{dP}{P} \\ &= [W_1(\alpha - \mu) + W_2(\beta - \mu)]dt + [W_1\sigma + W_2\gamma]dz \\ &\quad + [W_2\eta - (W_1 + W_2)\delta]dq, \end{aligned} \quad (27)$$

where $W_3 \equiv -(W_1 + W_2)$ has been substituted out.

⁴⁵ This assumption is much more acceptable than the usual homogeneous expectations. It is quite reasonable to expect that investors may have quite different estimates for current (and future) expected returns due to different levels of information, techniques of analysis, etc. However, most analysts calculate estimates of variances and covariances in the same way: namely, by using previous price data. Since all have access to the same price history, it is also reasonable to assume that their variance-covariance estimates may be the same.

⁴⁶ Itô's Lemma is the stochastic-analog to the fundamental theorem of the calculus because it states how to differentiate functions of Wiener processes. For a complete description and proof, see McKean [27]. A brief discussion can be found in Merton [33].

⁴⁷ See Merton [32] or [33].

Suppose a strategy, $W_j = W_j^*$, can be chosen such that the coefficients of dz and dq in (27) are always zero. Then, the dollar return on that portfolio, dY^* , would be nonstochastic. Since the portfolio requires zero investment, it must be that to avoid "arbitrage"⁴⁸ profits, the expected (and realized) return on the portfolio with this strategy is zero. The two portfolio and one equilibrium conditions can be written as a 3×2 linear system,

$$\begin{aligned}(\alpha - \mu)W_1^* + (\beta - \mu)W_2^* &= 0 \\ \sigma W_1^* + \gamma W_2^* &= 0 \\ -\delta W_1^* + (\eta - \delta)W_2^* &= 0.\end{aligned}\tag{28}$$

A nontrivial solution ($W_1^* \neq 0$; $W_2^* \neq 0$) to (28) exists if and only if

$$\frac{\beta - \mu}{\alpha - \mu} = \frac{\gamma}{\sigma} = \frac{\delta - \eta}{\delta}.\tag{29}$$

Because we make the "bucket shop" assumption, μ , α , δ , and σ are legitimate exogenous variables (relative to the option price), and β , γ , and η are to be determined so as to avoid dominance of any of the three securities. If (29) holds, then $\gamma/\sigma = 1 - \eta/\delta$, which implies from the definition of γ and in (26), that

$$\frac{SH_1}{H} = 1 - \frac{PH_2}{H}\tag{30}$$

or

$$H = SH_1 + PH_2.\tag{31}$$

Although it is not a sufficient condition, by Euler's theorem, (31) is a necessary condition for H to be first degree homogeneous in (S, P) as was conjectured in Section 2.

The second condition from (29) is that $\beta - \mu = \gamma(\alpha - \mu)/\sigma$, which implies from the definition of β and γ in (26) that

$$\begin{aligned}\frac{1}{2}\sigma^2 S^2 H_{11} + \rho\sigma\delta SP H_{12} + \frac{1}{2}\delta^2 P^2 H_{22} \\ + \alpha SH_1 + \mu PH_2 - H_3 - \mu H = SH_1(\alpha - \mu),\end{aligned}\tag{32}$$

or, by combining terms, that

$$\begin{aligned}\frac{1}{2}\sigma^2 S^2 H_{11} + \rho\sigma\delta SP H_{12} + \frac{1}{2}\delta^2 P^2 H_{22} + \mu SH_1 \\ + \mu PH_2 - H_3 - \mu H = 0.\end{aligned}\tag{33}$$

Substituting for H from (31) and combining terms, (33) can be rewritten as

$$\frac{1}{2}[\sigma^2 S^2 H_{11} + 2\rho\sigma\delta SP H_{12} + \delta^2 P^2 H_{22}] - H_3 = 0,\tag{34}$$

which is a second-order, linear partial differential equation of the parabolic type.

⁴⁸ "Arbitrage" is used in the qualified sense that the distributional and other assumptions are known to hold with certainty. A weaker form would say that if the return on the portfolio is nonzero, either the option or the common stock would be a dominated security. See Samuelson [44] or [45] for a discussion of this distinction.

If H is the price of a European warrant, then H must satisfy (34) subject to the boundary conditions:

$$H(0, P, \tau; E) = 0 \quad (34a)$$

$$H(S, 1, 0; E) = \text{Max}[0, S - E], \quad (34b)$$

since by construction, $P(0) = 1$.

Define the variable $x \equiv S/EP(\tau)$, which is the price per share of stock in units of exercise price-dollars payable at a *fixed date* in the future (the expiration date of the warrant). The variable x is a well-defined price for $\tau \geq 0$, and from (23), (24), and Itô's Lemma, the dynamics of x are described by the stochastic differential equation,

$$\frac{dx}{x} = [\alpha - \mu + \delta^2 - \rho\sigma\delta]dt + \sigma dz - \delta dq. \quad (35)$$

From (35), the expected return on x will be a function of S, P , etc., through α and μ , but the instantaneous variance of the return on x , $V^2(\tau)$, is equal to $\sigma^2 + \delta^2 - 2\rho\sigma\delta$, and will depend only on τ .

Motivated by the possible homogeneity properties of H , we try the change in variables, $h(x, \tau; E) \equiv H(S, P, \tau; E)/EP$ where h is assumed to be independent of P and is the warrant price evaluated in the same units as x . Substituting (h, x) for (H, S) in (34), (34a) and (34b), leads to the partial differential equation for h ,

$$\frac{1}{2}V^2x^2h_{11} - h_2 = 0, \quad (36)$$

subject to the boundary conditions, $h(0, \tau; E) = 0$, and $h(x, 0; E) = \text{Max}[0, x - 1]$. From inspection of (36) and its boundary conditions, h is only a function of x and τ , since V^2 is only a function of τ . Hence, the assumed homogeneity property of H is verified. Further, h does not depend on E , and so, H is actually homogeneous of degree one in $[S, EP(\tau)]$.

Consider a new time variable, $T \equiv \int_0^\tau V^2(s)ds$. Then, if we define $y(x, T) \equiv h(x, \tau)$ and substitute into (36), y must satisfy

$$\frac{1}{2}x^2y_{11} - y_2 = 0, \quad (37)$$

subject to the boundary conditions, $y(0, T) = 0$ and $y(x, 0) = \text{Max}[0, x - 1]$. Suppose we wrote the warrant price in its "full functional form," $H(S, P, \tau; E, \sigma^2, \delta^2, \rho)$. Then,

$$y = H(x, 1, T; 1, 1, 0, 0),$$

and is the price of a warrant with T years to expiration and exercise price of one dollar, on a stock with unit instantaneous variance of return, when the market rate of interest is zero over the life of the contract.

Once we solve (37) for the price of this "standard" warrant, we have, by a change of variables, the price for any European warrant. Namely,

$$H(S, P, \tau; E) = EP(\tau)y\left[S/EP(\tau), \int_0^\tau V^2(s)ds\right]. \quad (38)$$

Hence, for empirical testing or applications, one need only compute tables for the "standard" warrant price as a function of two variables, stock price and time to expiration, to be able to compute warrant prices in general.

To solve (37), we first put it in standard form by the change in variables $Z \equiv \log(x) + T/2$ and $\phi(Z, T) \equiv y(x, T)/x$, and then substitute in (37) to arrive at

$$0 = \frac{1}{2}\phi_{11} - \phi_2, \quad (39)$$

subject to the boundary conditions: $|\phi(Z, T)| \leq 1$ and $\phi(Z, 0) = \text{Max}[0, 1 - e^{-Z}]$. Equation (39) is a standard free-boundary problem to be solved by separation of variables or Fourier transforms.⁴⁹ Its solution is

$$y(x, T) = x\phi(Z, T) = [x\text{erfc}(h_1) - \text{erfc}(h_2)]/2, \quad (40)$$

where erfc is the error complement function which is tabulated, $h_1 \equiv -[\log x + \frac{1}{2}T]/\sqrt{2T}$, and $h_2 \equiv -[\log x - \frac{1}{2}T]/\sqrt{2T}$. Equation (40) is identical to (21) with $r = 0$, $\sigma^2 = 1$, and $E = 1$. Hence, (38) will be identical to (21) the B-S formula, in the special case of a non-stochastic and constant interest rate (i.e., $\delta = 0$, $\mu = r$, $P = e^{-r\tau}$, and $T \equiv \sigma^2\tau$).

Equation (37) corresponds exactly to Samuelson's equation⁵⁰ for the warrant price in his "linear" model when the stock price is log-normally distributed, with his parameters $\alpha = \beta = 0$, and $\sigma^2 = 1$. Hence, tables generated from (40) could be used with (38) for valuations of the Samuelson formula where $e^{-\alpha\tau}$ is substituted for $P(\tau)$ in (38).⁵¹ Since α in his theory is the expected rate of return on a risky security, one would expect that $e^{-\alpha\tau} < P(\tau)$. As a consequence of the following theorem, $e^{-\alpha\tau} < P(\tau)$ would imply that Samuelson's forecasted values for the warrants would be higher than those forecasted by B-S or the model presented here.

Theorem 14. For a given stock price, the warrant price is a non-increasing function of $P(\tau)$, and hence, a nondecreasing function of the τ -year interest rate.

Proof: it follows immediately, since an increase in P is equivalent to an increase in E which never increases the value of the warrant. Formally, H is a convex function of S and passes through the origin. Hence, $H - SH_1 \leq 0$. But from (31), $H - SH_1 = PH_2$, and since $P \geq 0$, $H_2 \leq 0$. By definition, $P(\tau)$ is a decreasing function of the τ -year interest rate. Q.E.D.

Because we applied only the terminal boundary condition to (34), the price function derived is for an European warrant. The correct boundary conditions for an American warrant would also include the arbitrage-boundary inequality

$$H(S, P, \tau; E) \geq \text{Max}[0, S - E]. \quad (34c)$$

Since it was assumed that no dividend payments or exercise price changes occur over the life of the contract, we know from Theorem 1, that if the formulation of this section is a "rational" theory, then

⁴⁹ For a separation of variables solution, see Churchill [9], pp. 154-156, and for the transform technique, see Dettman [12], p. 390. Also see McKean [26].

⁵⁰ In [42], p. 27.

⁵¹ The tables could also be used to evaluate warrants priced by the Sprenkle [49] formula. Warning: while the Samuelson interpretation of the " $\beta = \alpha$ " case implies that expected returns are equated on the warrant and the stock, the B-S interpretation does not. Namely, from [29], the expected return on the warrant satisfies $\beta = r + H_1S(\alpha - r)/H$, where H_1 can be computed from (21) by differentiation.

it will satisfy the stronger inequality $H \geq \text{Max}[0, S - EP(\tau)]$ [which is homogeneous in S and $EP(\tau)$], and the American warrant will have the same value as its European counterpart. Samuelson argued that solutions to equations like (21) and (38) will always have values at least as large as $\text{Max}[0, S - E]$, and Samuelson and Merton⁵² proved it under more general conditions. Hence, there is no need for formal verification here. Further, it can be shown that (38) satisfies all the theorems of Section 2.

As a direct result of the equal values of the European and American warrants, we have:

Theorem 15. The warrant price is a nondecreasing function of the variance of the stock price return.

Proof: from (38), the change in H with respect to a change in variance will be proportional to y_2 . But, y is the price of a legitimate American warrant and hence, must be a nondecreasing function of time to expiration, i.e., $y_2 \geq 0$. Q.E.D.

Actually, Theorem 15 is a special case of the general proposition (Theorem 8) proved in Section 2, that the more risky is the stock, the more valuable is the warrant. Although Rothschild and Stiglitz⁵³ have shown that, in general, increasing variance may not imply increasing risk, it is shown in Appendix 2 that variance is a valid measure of risk for this model.

We have derived the B-S warrant pricing formula rigorously under assumptions weaker than they postulate, and have extended the analysis to include the possibility of stochastic interest rates.

Because the original B-S derivation assumed constant interest rates in forming their hedge positions, it did not matter whether they borrowed or lent long or short maturities. The derivation here clearly demonstrates that the correct maturity to use in the hedge is the one which matches the maturity date of the option. "Correct" is used in the sense that if the price $P(\tau)$ remains fixed while the price of other maturities changes, the price of a τ -year option will remain unchanged.

The capital asset pricing model is a sufficient assumption to derive the formula. While the assumptions of this section are necessary for the intertemporal use of the capital asset pricing model,⁵⁴ they are not sufficient, e.g., we do not assume that interest rates are nonstochastic, that price dynamics are stationary, nor that investors have homogeneous expectations. All are required for the capital asset pricing model. Further, since we consider only the properties of three securities, we do not assume that the capital market is in full general equilibrium. Since the final formula is independent of α or μ , it will hold even if the observed stock or bond prices are transient, nonequilibrium prices.

The key to the derivation is that any one of the securities' returns over time can be perfectly replicated by continuous portfolio combinations of the other two. A complete analysis would require that

⁵² In [42] and [43], respectively.

⁵³ In [39].

⁵⁴ See Merton [31] for a discussion of necessary and sufficient conditions for a Sharpe-Lintner-Mossin type model to obtain in an intertemporal context. The sufficient conditions are rather restrictive.

all three securities' prices be solved for simultaneously which, in general, would require the examination of all other assets, knowledge of preferences, etc. However, because of "perfect substitutability" of the securities and the "bucket shop" assumption, supply effects can be neglected, and we can apply "partial equilibrium" analysis resulting in a "causal-type" formula for the option price as a function of the stock and bond prices.

This "perfect substitutability" of the common stock and borrowing for the warrant or the warrant and lending for the common stock explains why the formula is independent of the expected return on the common stock or preferences. The expected return on the stock and the investor's preferences will determine how much capital to invest (long or short) in a given company. The decision as to whether to take the position by buying warrants or by leveraging the stock depends only on their relative prices and the cost of borrowing. As B-S point out, the argument is similar to an intertemporal Modigliani-Miller theorem. The reason that the B-S assumption of the capital asset pricing model leads to the correct formula is that because it is an equilibrium model, it must necessarily rule out "sure-thing" profits among perfectly correlated securities, which is exactly condition (29). Careful study of both their derivations shows that (29) is the only part of the capital asset pricing model ever used.

The assumptions of this section are necessary for (38) and (40) to hold.⁵⁵ The continuous-trading assumption is necessary to establish perfect correlation among nonlinear functions which is required to form the "perfect hedge" portfolio mix. The Samuelson and Merton model⁵⁶ is an immediate counter-example to the validity of the formula for discrete-trading intervals.

The assumption of Itô processes for the assets' returns dynamics was necessary to apply Itô's Lemma. The further restriction that σ and δ be nonstochastic and independent of the price levels is required so that the option price change is due only to changes in the stock or bond prices, which was necessary to establish a perfect hedge and to establish the homogeneity property (31).⁵⁷ Clearly if investors did not agree on the value of $V^2(\tau)$, they would arrive at different values for the same warrant.

The B-S claim that (21) or (38) is the only formula consistent with capital market equilibrium is a bit too strong. It is not true that if the market prices options differently, then arbitrage profits are ensured. It is a "rational" option pricing theory relative to the assumptions of this section. If these assumptions held with certainty, then the B-S formula is the only one which all investors could agree on, and no deviant member could prove them wrong.⁵⁸

⁵⁵ If most of the "frictionless" market assumptions are dropped, it may be possible to show that, by substituting current institutional conditions, (38) and (40) will give lower bounds for the warrant's value.

⁵⁶ In [43].

⁵⁷ In the special case when interest rates are nonstochastic, the variance of the stock price return can be a function of the price level and the derivation still goes through. However, the resulting partial differential equation will not have a simple closed-form solution.

⁵⁸ This point is emphasized in a critique of Thorp and Kassouf's [53] "sure-thing" arbitrage techniques by Samuelson [45] and again, in Samuelson [44], footnote 6.

7. Extension of the model to include dividend payments and exercise price changes

■ To analyze the effect of dividends on unprotected warrants, it is helpful to assume a constant and known interest rate r . Under this assumption, $\delta = 0$, $\mu = r$, and $P(\tau) = e^{-r\tau}$. Condition (29) simplifies to

$$\beta - r = \gamma(\alpha - r)/\sigma. \quad (41)$$

Let $D(S, \tau)$ be the dividend per share unit time when the stock price is S and the warrant has τ years to expiration. If α is the instantaneous, *total* expected return as defined in (23), then the instantaneous expected return from price appreciation is $[\alpha - D(S, \tau)/S]$. Because $P(\tau)$ is no longer stochastic, we suppress it and write the warrant price function as $W(S, \tau; E)$. As was done in (25) and (26), we apply Itô's Lemma to derive the stochastic differential equation for the warrant price to be

$$\begin{aligned} dW &= W_1(dS - D(S, \tau)dt) + W_2d\tau + \frac{1}{2}W_{11}(dS)^2 \\ &= [\frac{1}{2}\sigma^2S^2W_{11} + (\alpha S - D)W_1 - W_2]dt + \sigma SW_1dz. \end{aligned} \quad (42)$$

Note: since the warrant owner is not entitled to any part of the dividend return, he only considers that part of the expected dollar return to the common stock due to price appreciation. From (42) and the definition of β and γ , we have that

$$\begin{aligned} \beta W &= \frac{1}{2}\sigma^2S^2W_{11} + (\alpha S - D)W_1 - W_2 \\ \gamma W &= \sigma SW_1. \end{aligned} \quad (43)$$

Applying (41) to (43), we arrive at the partial differential equation for the warrant price,

$$\frac{1}{2}\sigma^2S^2W_{11} + (rS - D)W_1 - W_2 - rW = 0, \quad (44)$$

subject to the boundary conditions, $W(0, \tau; E) = 0$, $W(S, 0; E) = \text{Max}[0, S - E]$ for a European warrant, and to the additional arbitrage boundary condition, $W(S, \tau; E) \geq \text{Max}[0, S - E]$ for an American warrant.

Equation (44) will not have a simple solution, even for the European warrant and relatively simple functional forms for D . In evaluating the American warrant in the "no-dividend" case ($D = 0$), the arbitrage boundary inequalities were not considered explicitly in arriving at a solution, because it was shown that the European warrant price never violated the inequality, and the American and European warrant prices were equal. For many dividend policies, the solution for the European warrant price will violate the inequality, and for those policies, there will be a positive probability of premature exercising of the American warrant. Hence, to obtain a correct value for the American warrant from (44), we must explicitly consider the boundary inequality, and transform it into a suitable form for solution.

If there exists a positive probability of premature exercising, then, for every τ , there exists a level of stock price, $C[\tau]$, such that for all $S > C[\tau]$, the warrant would be worth more exercised than if held. Since the value of an exercised warrant is always $(S - E)$, we have the appended boundary condition for (44),

$$W(C[\tau], \tau; E) = C[\tau] - E, \quad (44a)$$

where W satisfies (44) for $0 \leq S \leq C[\tau]$.

If $C[\tau]$ were a known function, then, after the appropriate change of variables, (44) with the European boundary conditions and (44a)

appended, would be a semiinfinite boundary value problem with a time-dependent boundary. However, $C[\tau]$ is not known, and must be determined as part of the solution. Therefore, an additional boundary condition is required for the problem to be well-posed.

Fortunately, the economics of the problem are sufficiently rich to provide this extra condition. Because the warrant holder is not contractually obliged to exercise his warrant prematurely, he chooses to do so only in his own best interest (i.e., when the warrant is worth more "dead" than "alive"). Hence, the only rational choice for $C[\tau]$ is that time-pattern which maximizes the value of the warrant. Let $f(S, \tau; E, C[\tau])$ be a solution to (44)–(44a) for a given $C[\tau]$ function. Then, the value of a τ -year American warrant will be

$$W(S, \tau; E) = \underset{[c]}{\text{Max}} f(S, \tau; E, C). \quad (45)$$

Further, the structure of the problem makes it clear that the optimal $C[\tau]$ will be independent of the current level of the stock price. In attacking this difficult problem, Samuelson⁵⁹ postulated that the extra condition was "high-contact" at the boundary, i.e.,

$$W_1(C[\tau], \tau; E) = 1. \quad (44b)$$

It can be shown⁶⁰ that (44b) is implied by the maximizing behavior described by (45). So the correct specification for the American warrant price is (44) with the European boundary conditions plus (44a) and (44b).

Samuelson and Samuelson and Merton⁶¹ have shown that for a proportional dividend policy where $D(S, \tau) = \rho S$, $\rho > 0$, there is always a positive probability of premature exercising, and hence, the arbitrage boundary condition will be binding for sufficiently large stock prices.⁶² With $D = \rho S$, (44) is mathematically identical to Samuelson's⁶³ "nonlinear" (" $\beta > \alpha$ ") case where his $\beta = r$ and his $\alpha = r - \rho$. Samuelson and McKean⁶⁴ analyze this problem in great detail. Although there are no simple closed-form solutions for finite-lived warrants, they did derive solutions for perpetual warrants which are power functions, tangent to the " $S - E$ " line at finite values of S .⁶⁵

⁵⁹ In [42].

⁶⁰ Let $f(x, c)$ be a differentiable function, concave in its second argument, for $0 \leq x \leq c$. Require that $f(c, c) = h(c)$, a differentiable function of c . Let $c = c^*$ be the c which maximizes f , i.e.,

$$f_2(x, c^*) = 0,$$

where subscripts denote partial derivatives. Consider the total derivative of f with respect to c along the boundary $x = c$. Then,

$$df/dc = dh/dc = f_1(c, c) + f_2(c, c).$$

For $c = c^*$, $f_2 = 0$. Hence, $f_1(c^*, c^*) = dh/dc$. In the case of the text, $h = c - E$, and the "high-contact" solution, $f_1(c^*, c^*) = 1$, is proved.

⁶¹ In [42] and [43], respectively.

⁶² For $D = \rho S$, the solution to (44) for the European warrant is

$$W = [e^{-\rho\tau} S \Phi(d_1) - E e^{-r\tau} \Phi(d_2)]$$

where Φ , d_1 , and d_2 are as defined in (21). For large S ,

$$W \sim [e^{-\rho\tau} S - E e^{-r\tau}]$$

which will be less than $(S - E)$ for large S and $\rho > 0$. Hence, the American warrant can be worth more "dead" than "alive."

⁶³ In [42].

⁶⁴ *Ibid.* In the appendix.

⁶⁵ *Ibid.*, p. 28.

A second example of a simple dividend policy is the constant one where $D = d$, a constant. Unlike the previous proportional policy, premature exercising may or may not occur, depending upon the values for d , r , E , and τ . In particular, a sufficient condition for no premature exercising was derived in Section 3. Namely,

$$E > \frac{d}{r}. \quad (13)$$

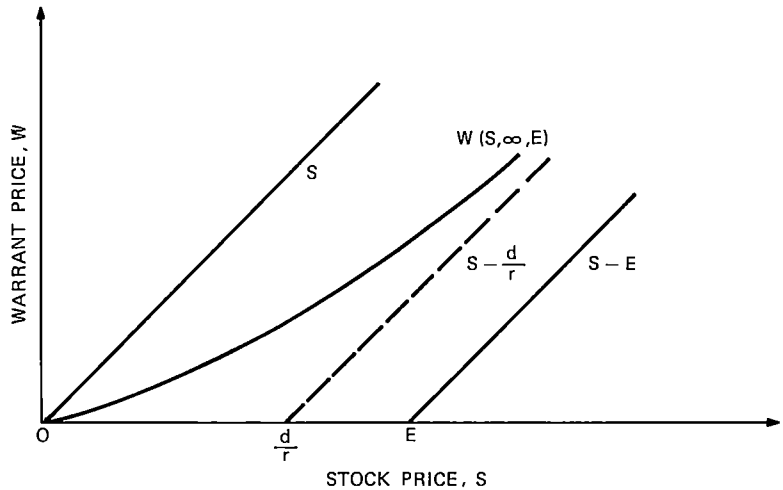
If (13) obtains, then the solution for the European warrant price will be the solution for the American warrant. Although a closed-form solution has not yet been found for finite τ , a solution for the perpetual warrant when $E > d/r$, is⁶⁶

$W(S, \infty; E)$

$$= S - \frac{d}{r} \left[1 - \frac{\left(\frac{2d}{\sigma^2 S}\right)^{2r/\sigma^2}}{\Gamma\left(2 + \frac{2r}{\sigma^2}\right)} M\left(\frac{2r}{\sigma^2}, 2 + \frac{2r}{\sigma^2}, \frac{-2d}{\sigma^2 S}\right) \right] \quad (46)$$

where M is the confluent hypergeometric function, and W is plotted in Figure 2.

FIGURE 2



⁶⁶ Make the change in variables: $Z \equiv \delta/S$ and

$$h(Z) \equiv \exp[Z]Z^{-\gamma}W$$

where

$$\delta \equiv 2d/\sigma^2$$

and

$$\gamma \equiv 2r/\sigma^2.$$

Then, substituting in (44), we have the differential equation for h :

$$Zh'' + (\gamma + 2 - Z)h' - 2h = 0,$$

whose general solution is $h = c_1 M(2, 2 + \gamma, Z) + c_2 Z^{-(\gamma+1)} M(1 - \gamma, -\gamma, Z)$ which becomes (46) when the boundary conditions are applied. Analysis of (46) shows that W passes through the origin, is convex, and is asymptotic to the line $(S - d/r)$ for large S , i.e., it approaches the common stock value less the present discounted value of all future dividends forgone by holding the warrant.

Consider the case of a continuously changing exercise price, $E(\tau)$, where E is assumed to be differentiable and a decreasing function of the length of time to maturity, i.e., $dE/d\tau = -dE/dt = -\dot{E} < 0$. The warrant price will satisfy (44) with $D = 0$, but subject to the boundary conditions,

$$W[S, 0; E(0)] = \text{Max}[0, S - E(0)]$$

and

$$W[S, \tau; E(\tau)] \geq \text{Max}[0, S - E(\tau)].$$

Make the change in variables $X \equiv S/E(\tau)$ and

$$F(X, \tau) \equiv W[S, \tau; E(\tau)]/E(\tau).$$

Then, F satisfies

$$\frac{1}{2}\sigma^2 X^2 F_{11} + \eta(\tau) X F_1 - \eta(\tau) F - F_2 = 0, \quad (47)$$

subject to $F(X, 0) = \text{Max}[0, X - 1]$ and $F(X, \tau) \geq \text{Max}[0, X - 1]$ where $\eta(\tau) \equiv r - \dot{E}/E$. Notice that the structure of (47) is identical to the pricing of a warrant with a fixed exercise price and a variable, but nonstochastic, "interest rate" $\eta(\tau)$. (I.e., substitute in the analysis of the previous section for $P(\tau)$, $\exp[-\int_0^\tau \eta(s)ds]$, except $\eta(\tau)$ can be negative for sufficiently large changes in exercise price.) We have already shown that for $\int_0^\tau \eta(s)ds \geq 0$, there will be no premature exercising of the warrant, and only the terminal exercise price should matter. Noting that $\int_0^\tau \eta(s)ds = \int_0^\tau [r + dE/d\tau]ds = r\tau + \log[E(\tau)/E(0)]$, formal substitution for $P(\tau)$ in (38) verifies that the value of the warrant is the same as for a warrant with a fixed exercise price, $E(0)$, and interest rate r . We also have agreement of the current model with (11) of Section 3, because $\int_0^\tau \eta(s)ds \geq 0$ implies $E(\tau) \geq E(0)\exp[-r\tau]$, which is a general sufficient condition for no premature exercising.

■ As the first example of an application of the model to other types of options, we now consider the rational pricing of the put option, relative to the assumptions in Section 7. In Section 4, it was demonstrated that the value of an European put option was completely determined once the value of the call option is known (Theorem 12). B-S give the solution for their model in equation (26). It was also demonstrated in Section 4 that the European valuation is not valid for the American put option because of the positive probability of premature exercising. If $G(S, \tau; E)$ is the rational put price, then, by the same technique used to derive (44) with $D = 0$, G satisfies

$$\frac{1}{2}\sigma^2 S^2 G_{11} + rSG_1 - rG - G_2 = 0, \quad (48)$$

subject to $G(\infty, \tau; E) = 0$, $G(S, 0; E) = \text{Max}[0, E - S]$, and $G(S, \tau; E) \geq \text{Max}[0, E - S]$.

From the analysis by Samuelson and McKean⁶⁷ on warrants, there is no closed-form solution to (48) for finite τ . However, using their techniques, it is possible to obtain a solution for the perpetual put option (i.e., $\tau = \infty$). For a sufficiently low stock price, it will be advantageous to exercise the put. Define C to be the largest value of the stock such that the put holder is better off exercising than continuing to hold it. For the perpetual put, (48) reduces to the ordinary

8. Valuing an American put option

⁶⁷ In [42].

differential equation,

$$\frac{1}{2}\sigma^2 S^2 G_{11} + rSG_1 - rG = 0, \quad (49)$$

which is valid for the range of stock prices $C \leq S \leq \infty$. The boundary conditions for (49) are:

$$G(\infty, \infty; E) = 0, \quad (49a)$$

$$G(C, \infty; E) = E - C, \text{ and} \quad (49b)$$

choose C so as to maximize the value of the option, which follows from the maximizing behavior arguments of the (49c) previous section.

From the theory of linear ordinary differential equations, solutions to (49) involve two constants, a_1 and a_2 . Boundary conditions (49a), (49b), and (49c) will determine these constants along with the unknown lower-bound, stock price, C . The general solution to (49) is

$$G(S, \infty; E) = a_1 S + a_2 S^{-\gamma}, \quad (50)$$

where $\gamma \equiv 2r/\sigma^2 > 0$. Equation (49a) requires that $a_1 = 0$, and (49b) requires that $a_2 = (E - C)C^\gamma$. Hence, as a function of C ,

$$G(S, \infty; E) = (E - C)(S/C)^{-\gamma}. \quad (51)$$

To determine C , we apply (49c) and choose that value of C which maximizes (51), i.e., choose $C = C^*$ such that $\partial G/\partial C = 0$. Solving this condition, we have that $C^* = \gamma E/(1 + \gamma)$, and the put option price is,

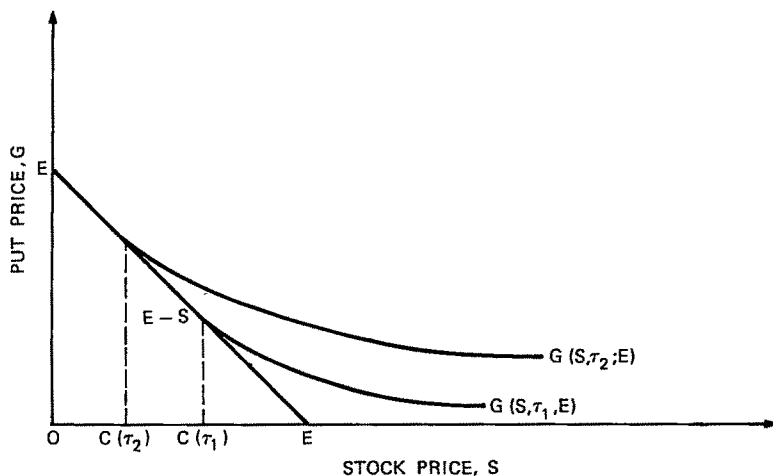
$$G(S, \infty; E) = \frac{E}{(1 + \gamma)} [(1 + \gamma)S/\gamma E]^{-\gamma}. \quad (52)$$

The Samuelson "high-contact" boundary condition

$$G_1(C^*, \infty; E) = -1,$$

as an alternative specification of boundary condition (49c), can be verified by differentiating (52) with respect to S and evaluating at $S = C^*$. Figure 3 illustrates the American put price as a function of the stock price and time to expiration.

FIGURE 3



■ As a second example of the application of the model to other types of options, we consider the rational pricing of a new type of call option called the "down-and-outer."⁶⁸ This option has the same terms with respect to exercise price, antidilution clauses, etc., as the standard call option, but with the additional feature that if the stock price falls below a stated level, the option contract is nullified, i.e., the option becomes worthless.⁶⁹ Typically, the "knock-out" price is a function of the time to expiration, increasing as the expiration date nears.

Let $f(S, \tau; E)$ be the value of an European "down-and-out" call option, and $B[\tau] = bE \exp[-\eta\tau]$ be the "knock-out" price as a function of time to expiration where it is assumed that $\eta \geq 0$ and $0 \leq b \leq 1$. Then f will satisfy the fundamental partial differential equation,

$$\frac{1}{2}\sigma^2 S^2 f_{11} + rSf_1 - rf - f_2 = 0, \quad (53)$$

subject to the boundary conditions,

$$\begin{aligned} f(B[\tau], \tau; E) &= 0 \\ f(S, 0; E) &= \text{Max}[0, S - E]. \end{aligned}$$

Note: if $B(\tau) = 0$, then (53) would be the equation for a standard European call option.

Make the change in variables, $x \equiv \log[S/B(\tau)]$; $T \equiv \sigma^2\tau$;

$$H(x, T) \equiv \exp[ax + \gamma T]f(S, \tau; E)/E,$$

and $a \equiv [r - \eta - \sigma^2/2]/\sigma^2$ and $\gamma \equiv r + a^2\sigma^2/2$. Then, by substituting into (53), we arrive at the equation for H ,

$$\frac{1}{2}H_{11} - H_2 = 0 \quad (54)$$

subject to

$$\begin{aligned} H(0, T) &= 0 \\ H(x, 0) &= e^{ax} \text{Max}[0, be^x - 1], \end{aligned}$$

which is a standard, semiinfinite boundary value problem to be solved by separation of variables or Fourier transforms.⁷⁰

Solving (54) and substituting back, we arrive at the solution for the "down-and-out" option,

$$\begin{aligned} f(S, \tau; E) &= [S \operatorname{erfc}(h_1) - Ee^{-r\tau} \operatorname{erfc}(h_2)]/2 \\ &\quad - (S/B[\tau])^{-\delta} [B[\tau] \operatorname{erfc}(h_3) - (S/B[\tau]) Ee^{-r\tau} \operatorname{erfc}(h_4)]/2, \end{aligned} \quad (55)$$

⁶⁸ See Snyder [48] for a complete description. A number of Wall Street houses are beginning to deal in this option. See *Fortune*, November, 1971, p. 213.

⁶⁹ In some versions of the "down-and-outer," the option owner receives a positive rebate, $R(\tau)$, if the stock price hits the "knock-out" price. Typically, $R(\tau)$ is an increasing function of the time until expiration [i.e., $R'(\tau) > 0$] with $R(0) = 0$. Let $g(S, \tau)$ satisfy (53) for $B(\tau) \leq S < \infty$, subject to the boundary conditions (a) $g(B[\tau], \tau) = R(\tau)$ and (b) $g(S, 0) = 0$. Then, $F(S, \tau; E) \equiv g(S, \tau) + f(S, \tau; E)$ will satisfy (53) subject to the boundary conditions (a) $F(B[\tau], \tau; E) = R(\tau)$ and (b) $F(S, 0; E) = \text{Max}[0, S - E]$. Hence, F is the value of a "down-and-out" call option with rebate payments $R(\tau)$, and $g(S, \tau)$ is the additional value for the rebate feature. See Dettman [12], p. 391, for a transform solution for $g(S, \tau)$.

⁷⁰ See Churchill [9], p. 152, for a separation of variables solution and Dettman [12], p. 391, for a transform solution.

where

$$\begin{aligned} h_1 &\equiv -[\log(S/E) + (r + \sigma^2/2)\tau]/\sqrt{2\sigma^2\tau}, \\ h_2 &\equiv -[\log(S/E) + (r - \sigma^2/2)\tau]/\sqrt{2\sigma^2\tau}, \\ h_3 &\equiv -[2\log(B[\tau]/E) - \log(S/E) + (r + \sigma^2/2)\tau]/\sqrt{2\sigma^2\tau}, \\ h_4 &\equiv -[2\log(B[\tau]/E) - \log(S/E) + (r - \sigma^2/2)\tau]/\sqrt{2\sigma^2\tau}, \end{aligned}$$

and $\delta \equiv 2(r - \eta)/\sigma^2$. Inspection of (55) and (21) reveals that the first bracketed set of terms in (55) is the value of a standard call option, and hence, the second bracket is the “discount” due to the “down-and-out” feature.

To gain a better perspective on the qualitative differences between the standard call option and the “down-and-out,” it is useful to go to the limit of a perpetual option where the “knock-out” price is constant (i.e., $\eta = 0$). In this case, (53) reduces to the ordinary differential equation

$$\frac{1}{2}\sigma^2 S^2 f'' + rSf' - rf = 0 \quad (56)$$

subject to

$$f(bE) = 0 \quad (56a)$$

$$f(S) \leq S, \quad (56b)$$

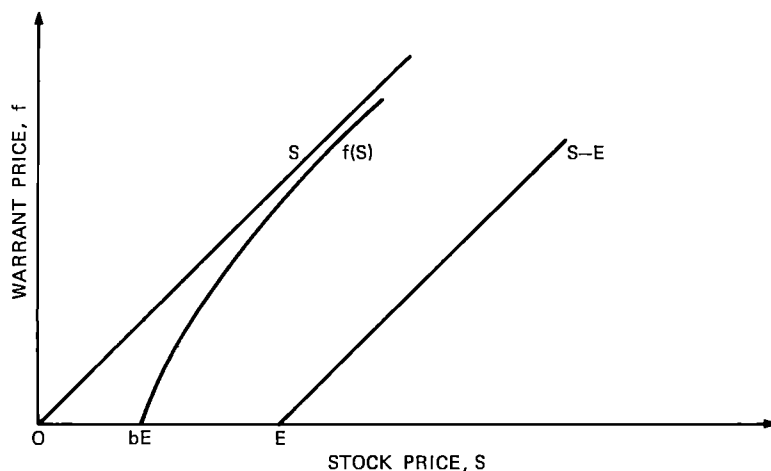
where primes denote derivatives and $f(S)$ is short for $f(S, \infty; E)$. By standard methods, we solve (56) to obtain

$$f(S) = S - bE(S/bE)^{-\gamma}, \quad (57)$$

where $\gamma \equiv 2r/\sigma^2$. Remembering that the value of a standard perpetual call option equals the value of the stock, we may interpret $bE(S/bE)^{-\gamma}$ as the “discount” for the “down-and-out” feature. Both (55) and (57) are homogeneous of degree one in (S, E) as are the standard options. Further, it is easy to show that $f(S) \geq \text{Max}[0, S - E]$, and although a tedious exercise, it also can be shown that $f(S, \tau; E) \geq \text{Max}[0, S - E]$. Hence, the option is worth more “alive” than “dead,” and therefore, (55) and (57) are the correct valuation functions for the American “down-and-out.”

From (57), the elasticity of the option price with respect to the stock price $[Sf'(S)/f(S)]$ is greater than one, and so it is a “levered” security. However, unlike the standard call option, it is a concave function of the stock price, as illustrated in Figure 4.

FIGURE 4



■ As our third and last example of an application of the model to other types of options, we consider the rational pricing of a callable American warrant. Although warrants are rarely issued as callable, this is an important example because the analysis is readily carried over to the valuation of other types of securities such as convertible bonds which are almost always issued as callable.

We assume the standard conditions for an American warrant except that the issuing company has the right to ("call") buy back the warrant at any time for a fixed price. Because the warrant is of the American type, in the event of a call, the warrant holder has the option of exercising his warrant rather than selling it back to the company at the call price. If this occurs, it is called "forced conversion," because the warrant holder is "forced" to exercise, if the value of the warrant exercised exceeds the call price.

The value of a callable warrant will be equal to the value of an equivalent noncallable warrant less some "discount." This discount will be the value of the call provision to the company. One can think of the callable warrant as the resultant of two transactions: the company sells a noncallable warrant to an investor and simultaneously, purchases from the investor an option to either "force" earlier conversion or to retire the issue at a fixed price.

Let $F(S, \tau; E)$ be the value of a callable American warrant; $H(S, \tau; E)$ the value of an equivalent noncallable warrant as obtained from equation (21), $C(S, \tau; E)$ the value of the call provision. Then $H = F + C$. F will satisfy the fundamental partial differential equation,

$$\frac{1}{2}\sigma^2 S^2 F_{11} + rSF_1 - rF - F_2 = 0 \quad (58)$$

for $0 \leq S \leq \bar{S}$ and subject to

$$\begin{aligned} F(0, \tau; E) &= 0, \\ F(S, 0; E) &= \text{Max}[0, S - E] \\ F(\bar{S}, \tau; E) &= \text{Max}[K, \bar{S} - E], \end{aligned}$$

where K is the call price and $\bar{S}$ is the (yet to be determined) level of the stock price where the company will call the warrant. Unlike the case of "voluntary" conversion of the warrant because of unfavorable dividend protection analyzed in Section 7, $\bar{S}$ is not the choice of the warrant owner, but of the company, and hence will not be selected to maximize the value of the warrant.

Because $C = H - F$ and H and F satisfy (58), C will satisfy (58) subject to the boundary conditions,

$$\begin{aligned} C(0, \tau; E) &= 0 \\ C(S, 0; E) &= 0 \\ C(\bar{S}, \tau; E) &= H(\bar{S}, \tau; E) - \text{Max}[K, \bar{S} - E]. \end{aligned}$$

Because $\bar{S}$ is the company's choice, we append the maximizing condition that $\bar{S}$ be chosen so as to maximize $C(S, \tau; E)$ making (58) a well-posed problem. Since $C = H - F$ and H is not a function of $\bar{S}$, the maximizing condition on C can be rewritten as a minimizing condition on F .

In general, it will not be possible to obtain a closed-form solution to (58). However, a solution can be found for the perpetual warrant. In this case, we know that $H(S, \tau; E) = S$, and (58) reduces to the

ordinary differential equation

$$\frac{1}{2}\sigma^2 S^2 C'' + rSC' - rC = 0 \quad (59)$$

for $0 \leq S \leq \bar{S}$ and subject to

$$C(0) = 0$$

$$C(\bar{S}) = \bar{S} - \text{Max}(K, \bar{S} - E)$$

Choose $\bar{S}$ so as to maximize C ,

where $C(S)$ is short for $C(S, \infty; E)$ and primes denote derivatives. Solving (59) and applying the first two conditions, we have

$$C(S) = (1 - \text{Max}[K/\bar{S}, 1 - E/\bar{S}])S. \quad (60)$$

Although we cannot apply the simple calculus technique for finding the maximizing $\bar{S}$, it is obviously $\bar{S} = K + E$, since for $\bar{S} < K + E$, C is an increasing function of $\bar{S}$ and for $\bar{S} > K + E$, it is a decreasing function. Hence, the value of the call provision is

$$C(S) = \left(\frac{E}{K + E} \right) S, \quad (61)$$

and because $F = H - C$, the value of the callable perpetual warrant is

$$F(S) = \left(\frac{K}{K + E} \right) S. \quad (62)$$

11. Conclusion

■ It has been shown that a B-S type model can be derived from weaker assumptions than in their original formulation. The main attractions of the model are: (1) the derivation is based on the relatively weak condition of avoiding dominance; (2) the final formula is a function of "observable" variables; and (3) the model can be extended in a straightforward fashion to determine the rational price of any type option.

The model has been applied with some success to empirical investigations of the option market by Black and Scholes and to warrants by Leonard.⁷¹

As suggested by Black and Scholes and Merton,⁷² the model can be used to price the various elements of the firm's capital structure. Essentially, under conditions when the Modigliani-Miller theorem obtains, we can use the total value of the firm as a "basic" security (replacing the common stock in the formulation of this paper) and the individual securities within the capital structure (e.g., debt, convertible bonds, common stock, etc.) can be viewed as "options" or "contingent claims" on the firm and priced accordingly. So, for example, one can derive in a systematic fashion a risk-structure of interest rates as a function of the debt-equity ratio, the risk-class of the firm, and the riskless (in terms of default) debt rates.

Using the techniques developed here, it should be possible to develop a theory of the term structure of interest rates along the

⁷¹ In [5] and [25], respectively.

⁷² In [4] and [29], respectively.

lines of Cootner and Merton.⁷³ The approach would also have application in the theory of speculative markets.

Appendix 1⁷⁴

■ Theorems 9 and 10 state that warrants whose common stock per dollar returns possess *distributions* that are independent of *stock price* levels (henceforth, referred to as D.I.S.P.) are: (1) homogeneous of degree one in stock price S and exercise price E —Theorem 9 and (2) convex in S —Theorem 10. This appendix exhibits via counterexample the insufficiency of the posited assumptions *sans* D.I.S.P. for the proof of Theorems 9 and 10.

First, we posit a very simple, noncontroversial, one-period European warrant pricing function, W :

$$W(S, \lambda) = K \int_{E/S}^{\infty} (S\hat{Z} - E) dP(\hat{Z}; S, \lambda), \quad (A1)$$

wherein: $1 > K > 0$ is a discounting factor which is deemed (somewhat erroneously) to be constant at this point in time (i.e., independent of S),

$\lambda \in [0, 1]$ is a parameter of the distribution, dP ,

$$\hat{Z} \equiv Z + \lambda g(S)\epsilon \equiv Z + U(S, \lambda) \equiv \text{Common stock per dollar return}, \quad (A2)$$

Z and ϵ are independent random variables such that $E(\epsilon|Z) = 0$.

The function $g(S)$ has the following properties for our example: $g(S)\epsilon \in (0, 1)$, $\frac{dg(S)}{ds} < 0$, $dP(\hat{Z}; S, \lambda)$ is the Stieltjes integral representation of the probability density which is equivalent to the convolution of the probability densities of Z and U .

In constructing the counterexample, we choose the following uniform distributions for Z and U :

$$f(\epsilon) = (1/2) \quad \text{for} \quad -1 \leq \epsilon \leq 1 \quad (A3)$$

$$= 0 \quad \text{elsewhere}$$

$$\rightarrow f(U) = \frac{1}{2\lambda g(S)} \quad \text{for} \quad -\lambda g(S) \leq U \leq \lambda g(S)$$

$$= 0 \quad \text{elsewhere}$$

$$h(Z) = (1/2) \quad \text{for} \quad 1 \leq Z \leq 3$$

$$= 0 \quad \text{elsewhere.} \quad (A4)$$

The convoluted density would then be:

$$\frac{dP}{d\hat{Z}}(\hat{Z}; S, \lambda) = \frac{\hat{Z} - 1 + \lambda g(S)}{4\lambda g(S)} \quad \text{for} \quad 1 - \lambda g(S) \leq \hat{Z} \leq 1 + \lambda g(S) \quad (A5)$$

⁷³ In [11] and [29], respectively.

⁷⁴ I thank B. Goldman of M.I.T. for constructing this example and writing the appendix.

$$\begin{aligned}
&= (1/2) \quad \text{for } 1 + \lambda g(S) \leq \hat{Z} \leq 3 - \lambda g(S) \\
&= \frac{3 + \lambda g(S) - \hat{Z}}{4\lambda g(S)} \\
&\quad \text{for } 3 - \lambda g(S) \leq \hat{Z} \leq 3 + \lambda g(S) \\
&= 0 \text{ elsewhere.}
\end{aligned}$$

As a further convenience, we choose the exercise price, E , to be in the neighborhood of twice the stock price, S , and evaluate (A1):

$$W(S, \lambda) = K[E^2/4S - 3E/2 + 9S/4 + \lambda^2 g(S)^2 S/12]. \quad (\text{A6})$$

By inspection of (A6), we notice that W is not homogeneous of degree one in S and E . Moreover, the convexity of W can be violated (locally) (i.e., $\frac{d^2 W}{dS^2}$ can become negative) by choosing a sufficiently negative $\frac{d^2 g(S)}{dS^2}$:

$$\begin{aligned}
\frac{d^2 W}{dS^2} = & \\
& K \left(E^2/2S^3 + \lambda^2/6 \left[2g(S)dg/ds + \frac{S(dg)^2}{(dS)} + Sg(S) \frac{d^2 g(S)}{dS^2} \right] \right) \stackrel{(\text{A7})}{\geq} 0.
\end{aligned}$$

Thus, our example has shown Theorems 9 and 10 to be not generally consistent with a non-D.I.S.P. environment; however, we can verify Theorems 9 and 10 for the D.I.S.P. subcase of our example, since by construction setting $\lambda = 0$ reinstates the D.I.S.P. character of the probability distribution. By inspection, we observe that when $\lambda = 0$, the right-hand side of (A6) is homogeneous of degree one in S and E , while the right-hand side of (A7) is $KE^2/2S^3 > 0$, verifying the convexity theorem.

Appendix 2

■ It was stated in the text that Theorem 15 is really a special case of Theorem 8, i.e., variance is a consistent measure of risk in the B-S model. To prove consistency, we use the equivalent, alternative definition (Rothschild and Stiglitz)⁷⁶ of more risky that X is more risky than Y if $E[X] = E[Y]$ and $EU(X) \leq EU(Y)$ for every concave function U .

Since the B-S formula for warrant price, (21), is independent of the expected return on the stock and since the stock returns are assumed to be log normally distributed, different securities are distinguished by the single parameter, σ^2 . Therefore, without loss of generality, we can assume that $\alpha = 0$, and prove the result by showing that for every concave U , $EU(Z)$ is a decreasing function of σ , where Z is a log-normal variate with $E[Z] = 1$ and the variance of

$\log(Z)$ equal to σ^2 :

$$EU(Z) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^\infty U(Z) \exp\left\{-[\log Z + (1/2)\sigma^2]^2/2\sigma^2\right\} dZ/Z$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty U(e^{\sigma x - (1/2)\sigma^2}) e^{-(1/2)x^2} dx,$$

for $x \equiv [\log Z + (1/2)\sigma^2]/\sigma$;

$$\partial EU(Z)/\partial \sigma = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty U'(\) \exp[-(1/2)(x - \sigma)^2](x - \sigma) dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty U'(e^{\sigma y + (1/2)\sigma^2}) y e^{-1/2 y^2} dy, \text{ for } y \equiv x - \sigma$$

$$\equiv \text{Covariance } [U'(e^{\sigma y + (1/2)\sigma^2}), y].$$

But, $U'(\)$ is a decreasing function of y by the concavity of U . Hence, by Theorem 236, Hardy et al.,⁷⁶ $\text{Cov}[U', y] < 0$. Therefore, $\partial EU/\partial \sigma < 0$ for all concave U

⁷⁶ In [16], p. 168

References

1. AYRES, R. F. "Risk Aversion in the Warrant Markets." *Industrial Management Review*, Vol. 50, No. 1 (Fall 1963), pp. 45-53; reprinted in Cootner [10], pp. 497-505.
2. BACHELIER, L. *Theory of Speculation* (translation of 1900 French edition), in Cootner [10], pp. 17-78.
3. BAUMOL, W. J., MALKIEL, B. G., AND QUANDT, R. E. "The Valuation of Convertible Securities." *Quarterly Journal of Economics*, Vol. 80, No. 1 (February 1966), pp. 48-59.
4. BLACK, F. AND SCHOLES, M. "The Pricing of Options and Corporate Liabilities," forthcoming in *Journal of Political Economy*.
5. ———. "The Valuation of Option Contracts and a Test of Market Efficiency." *Journal of Finance*, Vol. 27, No. 2 (May 1972).
6. ———. "Some Evidence on the Profitability of Trading in Put and Call Options," in Cootner [10], pp. 475-496.
7. BONESS, A. J. "Elements of a Theory of Stock-Option Value," *Journal of Political Economy*, Vol. 72, No. 2 (April 1964), pp. 163-175.
8. CHEN, A. H. Y. "A Model of Warrant Pricing in a Dynamic Market," *Journal of Finance*, Vol. 25, No. 5 (December 1970).
9. CHURCHILL, R. V. *Fourier Series and Boundary Value Problems*. 2nd ed. New York: McGraw-Hill, 1963.
10. COOTNER, P. H., ed. *The Random Character of Stock Market Prices*. Cambridge: M.I.T. Press, 1964.
11. ———. "The Stochastic Theory of Bond Prices." Mimeographed. Massachusetts Institute of Technology, December 1966.
12. DETTMAN, J. W. *Mathematical Methods in Physics and Engineering*. 2nd ed. New York: McGraw-Hill, 1969.
13. FAMA, E. F. "Efficient Capital Markets: A Review of Theory and Empirical Work." *Journal of Finance*, Vol. 25, No. 2 (May 1970).
14. GIGUERE, G. "Warrants: A Mathematical Method of Evaluation." *Analysts Journal*, Vol. 14, No. 5 (November 1958), pp. 17-25.
15. HALLINGBY, P., JR. "Speculative Opportunities in Stock Purchase Warrants." *Analysts Journal*, Vol. 3, No. 3 (1947).
16. HARDY, G. H., LITTLEWOOD, J. E. AND PÖLYA, G. *Inequalities*. Cambridge: The University Press, 1959.

17. HAUSMAN, W. H. AND WHITE, W. L. "Theory of Option Strategy under Risk Aversion." *Journal of Financial and Quantitative Analysis*, Vol. 3, No. 3 (September 1968).
18. KASSOUF, S. T. *Evaluation of Convertible Securities*. Maspeth, N. Y.: Analytic Investors Inc., 1962.
19. ———. "Stock Price Random Walks: Some Supporting Evidence." *Review of Economics and Statistics*, Vol. 50, No. 2 (May 1968), pp. 275–278.
20. ———. *A Theory and an Econometric Model for Common Stock Purchase Warrants*. Ph.D. dissertation, Columbia University. New York: Analytical Publishers Co., 1965.
21. KRUIZENGA, R. J. "Introduction to the Option Contract," in Cootner [10], pp. 277–391.
22. ———. "Profit Returns from Purchasing Puts and Calls," in Cootner [10], pp. 392–411.
23. ———. *Put and Call Options: A Theoretical and Market Analysis*. Unpublished Ph.D. dissertation, M.I.T., 1956.
24. KUSHNER, H. J. *Stochastic Stability and Control*. New York: Academic Press, 1967.
25. LEONARD, R. J. "An Empirical Examination of a New General Equilibrium Model for Warrant Pricing." Unpublished M.S. thesis, M.I.T., September 1971.
26. MCKEAN, H. P., JR. "Appendix: A Free Boundary Problem for the Heat Equation Arising from a Problem in Mathematical Economics." *Industrial Management Review*, Vol. 6, No. 2 (Spring 1965), pp. 32–39; reprinted in [40], Chapter 199.
27. MERTON, R. C. *Stochastic Integrals*. New York: Academic Press, 1969.
28. ———. "Appendix: Continuous-Time Speculative Processes." (Appendix to Samuelson [45]), in *Mathematical Topics in Economic Theory and Computation*, SIAM, Philadelphia, 1972.
29. ———. "A Dynamic General Equilibrium Model of the Asset Market and its Application to the Pricing of the Capital Structure of the Firm." Sloan School of Management Working Paper #497–70, M.I.T., (December 1970).
30. ———. "An Empirical Investigation of the Samuelson Rational Warrant Pricing Theory," Chapter V in *Analytical Optimal Control Theory as Applied to Stochastic and Non-Stochastic Economics*, unpublished Ph.D. dissertation, M.I.T., 1970.
31. ———. "An Intertemporal Capital Asset Pricing Model," forthcoming in *Econometrica*.
32. ———. "Lifetime Portfolio Selection under Uncertainty: The Continuous-Time Case," *Review of Economics and Statistics*, Vol. 51, No. 3 (August 1969).
33. ———. "Optimum Consumption and Portfolio Rules in a Continuous-Time Model." *Journal of Economic Theory*, Vol. 3, No. 4 (December 1971).
34. ———. "Restrictions on Rational Option Pricing: A Set of Arbitrage Conditions." Mimeographed. Massachusetts Institute of Technology, August 1968.
35. MILLER, M. AND MODIGLIANI, F. "Dividend Policy, Growth, and the Valuation of Shares." *Journal of Business*, Vol. 34, No. 4 (October 1961).
36. MORRISON, R. J. "The Warrants or the Stock?" *Analysts Journal*, Vol. 13, No. 5 (November 1957).
37. PEASE, F. "The Warrant—Its Powers and Its Hazards." *Financial Analysts Journal*, Vol. 19, No. 1 (January–February 1963).
38. PLUM, V. L. AND MARTIN, T. J. "The Significance of Conversion Parity in Valuing Common Stock Warrants." *The Financial Review* (February 1966).
39. ROTHSCHILD, M. AND STIGLITZ, J. E. "Increasing Risk: I. A Definition." *Journal of Economic Theory*, Vol. 2, No. 3 (September 1970).
40. SAMUELSON, P. A. *The Collected Scientific Papers of Paul A. Samuelson*. Vol. 3. R. C. Merton, ed. Cambridge: M.I.T. Press, 1972.
41. ———. "Proof That Properly Anticipated Prices Fluctuate Randomly." *Industrial Management Review*, Vol. 6, No. 2 (Spring 1965), pp. 41–50; reprinted in [40], Chapter 198.
42. ———. "Rational Theory of Warrant Pricing." *Industrial Management Review*, Vol. 6, No. 2 (Spring 1965), pp. 13–31; reprinted in [40], Chapter 199.
43. ——— AND MERTON, R. C. "A Complete Model of Warrant Pricing That Maximizes Utility." *Industrial Management Review*, Vol. 10, No. 2 (Winter 1969), pp. 17–46; reprinted in [40], Chapter 200.

44. ———. "Mathematics of Speculative Price," in *Mathematical Topics in Economic Theory and Computation*, SIAM, Philadelphia, 1972.
45. ———. Review of [53]. *Journal of American Statistical Association*, Vol. 63, No. 323 (September 1968), pp. 1049–1051.
46. SHELTON, J. P. "The Relation of the Pricing of a Warrant to the Price of Its Associated Stock." *Financial Analysts Journal*, Vol. 23, Nos. 3–4 (Part I: May-June 1967) and (Part II: July-August 1967).
47. SLATER, L. J. "Confluent Hypergeometric Functions," Chapter 13 in *Handbook of Mathematical Functions*. National Bureau of Standards, Applied Mathematics Series, 55, August, 1966.
48. SNYDER, G. L., "Alternative Forms of Options." *Financial Analysts Journal*, Vol. 25, No. 1 (September–October 1969), pp. 93–99.
49. SPRENKLE, C. M. "Warrant Prices as Indicators of Expectations and Preferences." *Yale Economic Essays* I, pp. 172–231; reprinted in Cootner [10], pp. 412–474.
50. STIGLITZ, J. E. "On Some Aspects of the Pure Theory of Corporate Finance: Bankruptcies and Take-Overs." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 2 (Autumn 1972), pp. 458–482.
51. ———. "A Re-Examination of the Modigliani-Miller Theorem." *The American Economic Review*, Vol. 59, No. 5 (December 1969).
52. STOLL, H. R. "The Relationship between Put and Call Option Prices." *Journal of Finance*, Vol. 24, No. 4 (December 1969), pp. 802–824.
53. THORP, E. O. AND KASSOUF, S. T. *Beat the Market*. New York: Random House, 1967.
54. VAN HORNE, J. C. "Warrant Valuation in Relation to Volatility and Opportunity Costs." *Industrial Management Review*, Vol. 10, No. 3 (Spring 1969), pp. 17–32.

Predicting railroad bankruptcies in America

Edward I. Altman

Associate Professor of Finance
New York University

The purpose of this paper is to discuss the urgent need for an early-warning system covering the historically failure-prone railroad industry and to develop a tool for providing such a system. A multivariate statistical technique called linear discriminant analysis is utilized to identify and quantify those financial measures which are effective indicators of bankruptcies. A model which combined several financial statement ratios proved to be extremely accurate in predicting railroad bankruptcies at one and two annual financial statement dates prior to failure. Subsequent tests on additional railroad samples confirm the validity of the model. Finally, currently existing railroads in America are assessed for their bankruptcy potential by this diagnostic model.

1. Introduction

■ During the recession of 1970 and in the subsequent year, five of the nation's major railroads petitioned the federal courts for bankruptcy-reorganization under the provisions of the National Bankruptcy Act. These roads accounted for over \$5 billion in corporate assets, the vast majority of which are included in the giant Penn Central (P-C) railroad complex. The P-C bankruptcy, indeed, was the largest corporate failure in history. The event of bankruptcy, however, represented only a final phase of a sequence in which P-C and its predecessor companies displayed extraordinary resilience in a long history of solvency. And this resilience is all the more extraordinary when one considers that one of the pervasive characteristics of America's railroad industry is its propensity for failure. One analyst recently estimated that the number of railroad receiverships, bankruptcies, and liquidations during the period 1876-1970 exceeds 1,100.¹

Dr. Altman holds the B.A. degree from the City College of New York and the M.B.A. and Ph.D. from U.C.L.A. (1967). He is an Associate Professor of Finance at New York University, Graduate School of Business. Professor Altman is an economic advisor to the Commission on the Bankruptcy Laws of the United States. His book, *Corporate Bankruptcy in America* (Lexington, Mass.: Heath-Lexington, 1971), explores the bankruptcy phenomenon in depth. Professor Altman's current research interests involve capital market tests in France, credit scoring models, and investment bond quality ratings.

The author acknowledges the helpful assistance of Professors J. Fred Weston (U.C.L.A.) and Ernest Bloch (N.Y.U.) and the research assistance of Kathrine Alamo, Michael Berson, and Dakshinamurthy Dasari. Special mention is also given to the Reviewer and Editors of *The Bell Journal*. Any remaining errors are of course the province of the author. An earlier version of this paper is *Working Paper # 34*, New York University, School of Business. Part of this paper was written at the Centre d'Enseignement Supérieur des Affaires in Jouy-en-Josas, France, during a leave of absence, 1971-1972.

¹ See Vance [26]. This estimate compares with an Interstate Commerce Commission (ICC) estimate of 840 failures between 1894-1942.

More than any other industry in the business history of the United States, railroads have continually experienced operating and financial problems. One wonders how this industry with such enormous "public-interest" qualities, beneficiary of governmental subsidies, and supervision by the ICC since 1887, could continue to be plagued by failures without some decisive measures taken to reduce or mitigate them. At the very least, one would expect that some diagnostic work would have been undertaken to predict failure as a means of forestalling that melancholy event.

Indeed, it shall be the purpose of this paper to attempt to fill that gap by developing a statistical model which can be utilized to predict the occurrence of railroad bankruptcies. Two previous studies by this writer were concerned with a general prediction model, based on published financial data, to detect future bankruptcies among manufacturing firms,² and more recently to investigate the efficacy of building a similar model for railroads.³ The former paper utilized a multiple discriminant statistical model to accurately predict bankruptcy up to two annual financial statement dates prior to failure, while the latter discussed the qualities of railroads which seemed compatible with a predictive model. The present paper develops and tests the latter model. We conclude that the empirical results appear to confirm the model's usefulness.

The paper begins with a brief discussion of the need for such a predictive model, and presents some of the intractable problems that beset the industry (Section 2). In Section 3 we describe the statistical technique to be utilized and review the relevant characteristics of the railroad industry. Section 4 presents the development of the model with the empirical results. An original sample of twenty-one bankrupt railroads and industry averages are used to construct the parameters of our model, and several validation tests are conducted. An alternate strategy of comparing bankrupt roads to nonbankrupt ones is also examined. The model is then applied to the currently existing (as of the end of 1970) major railroads in the United States in order to explore the state of the industry and its members. Finally, Section 5 discusses the paper's implications.

■ A great deal of publicity has been given to the recent upsurge in railroad problems and to the number of actual and prospective bankruptcies. At the end of 1970, 16 percent of the nation's railroad assets were in the hands of receivers or trustees. In the 1970-1971 recession the industry's problems appeared to be one-half as serious as in the depression years of the 1930s, when approximately one-third of railroad trackage was in receivership. But they are comparable to problems in the post 1893-1897 depression when one-sixth of the mileage and one-quarter of the total capitalization were in receivership. In all, the latter period had about 90 different failure roads. The fact that the absolute *number* of bankrupt railroads today is relatively small should not deceive the analyst, since by consolidations, mergers, etc. the number of Class I railroads (\$5 million or more in revenues) in America has fallen dramatically.

2. Railroad problems

² See Altman [2].

³ See Altman [3].

The present dire state of affairs has prompted the present ICC Chairman to call for appropriate measures to combat bankruptcies and in fact to develop procedures for spotlighting potential failure situations.⁴ It is precisely in the area of an “early-warning” device that this paper places its major emphasis. Unfortunately, we have not observed further evidence of such endeavors by the ICC.

Railroad bankruptcies have enormous consequences to the public besides the obvious costs to corporate suppliers, creditors, and stockholders. The social costs due to reduced service, higher rates, and government subsidies and/or loan guarantees necessitating higher taxes shift the burden of failure among economic units in ways that no cost-benefit analysis could justify. And this unfortunate state of affairs appears to be continuing without abatement.

The large number of railroad failures has evolved due to a chain of unfavorable circumstances and poor management on a micro-economic level and also due to the regulatory nature of the industry. Because of the numerous legislative moves to regulate the transportation industry, from the Interstate Commerce Act (1887) to the Department of Transportation Act (1966), we now have an example of government-regulated cartelization which has been characterized by an inability to adjust to change and, at times, by grossly inefficient management. Rather than review the history of railroad “progress” in the U. S., it will serve the purposes of our discussion to summarize the major problem areas.⁵

The major reasons for the railroad industry’s dismal performances are: (1) its inability to meet competitive conditions due to an inflexible pricing and cost structure; (2) its acute susceptibility to large net income losses during periods of economic stress due to a fixed asset and liability structure which is heavily leveraged; (3) its excess capacity; (4) the acute labor and manpower rigidities; and (5) a shortage of innovative management. Many of these problems are the by-products of government regulation and industry rigidities. These have produced an industry pricing structure dictated, in part, by political and noneconomic terms rather than comparative advantage considerations.⁶ While it is not the intent of this discussion to suggest changes in the railroad industry, it should be pointed out that several noted economists favor deregulation resulting in allegedly desirable shifts in resource allocation in our transportation industries.⁷

The railroad industry is a classic illustration of the magnification effect caused by high operating and financial leverage structures. The leverage concept affirms that during periods of economic downturns those firms which are more heavily leveraged will experience greater swings in returns on equity (possibly to negative levels) and

⁴ In the aftermath of the Penn Central bankruptcy, ICC Chairman George Stafford commented on expected efforts: “Mr. Stafford said that beyond the Penn Central investigation, the ICC probably will develop a number of rule-making procedures aimed at fashioning guidelines for providing ‘early warnings’ of impending bankruptcies of other roads” [18].

⁵ An excellent review of the industry’s development and the resulting problems can be derived from the following sources among others: Bray [8]; Hilton [17]; Meyer, Peck, Stenason, and Zwick [20]; Peck [22], pp. 340–365; and Weston [28].

⁶ See J. Meyer et al. [20], p. 8.

⁷ See Weston [28], pp. 317–324 and J. Meyer et al. [20], p. 258.

be more vulnerable to failure than low fixed-cost firms. Combine this situation with recurring liquidity problems common to railroads, and the reasons for the high incidence of railroad bankruptcies become clear. In addition, the industry appears to be quite vulnerable to general economic declines and credit crunches such that the induced effect of the business cycle on railroads is sizeable.

The railroad industry is quite vulnerable to business cycle fluctuations, but unlike other cyclical industries, the asset and capital structures do not reflect this high risk situation. In Section 4 of this paper we observe that for those roads which went bankrupt, the percentage of total capital contributed by creditors one annual statement prior to bankruptcy averaged 91.2 percent, while the industry average was approximately 50 percent. It is questionable whether many firms could withstand even small declines in revenues in this extremely vulnerable condition of financial leverage. This implies a chronic rather than an episodic industry problem. The high risk situation, in the light of historically low returns on invested capital, provides valid grounds for criticism of the industry's overall managerial efficiency.

The case for railroad overcapacity has been made on economic grounds many times with periodic reports advocating railroad mergers and abandonments. In fact, however, government sanctions on mergers have often caused certain facilities to be retained for which abandonment was justifiable on economic grounds. Since the railroad industry has been marked by parallel and duplicative facilities, a strong case could be made for operating economies from mergers. Ironically, it was the merger of the two largest railroads in 1968 which helped plunge the merged Penn Central Transportation Company into bankruptcy. With this merger failure and the subsequent cancellation of merger discussions between the Chesapeake and Ohio and Norfolk and Western Railways (March 20, 1971) some analysts expect that the trend toward railroad merger is temporarily and possibly permanently derailed.⁸

The industry's labor situation is critical. Although we do not take a definitive position on this controversial topic, it is fair to point out that union featherbedding and rigid work rules are often cited as contributions to the railroad's problems. While some analysts testify that labor costs are the industry's number one problem—one of the primary causes of the Penn Central collapse and the failure of the unsuccessful merger⁹—others protest that the labor problems

⁸ Two acknowledged railroad experts, Isabel Benham and Arthur Jansen, attested to the probable disappearance of railroad mergers, except in isolated cases, in a recent interview [5] Mr. Jansen views the Penn Central bankruptcy as the industry "savior" because of the favorable legislation expected in its wake and the ridding of unprofitable passenger operations. Miss Benham, on the other hand, views the railroads in 1971 as being at the crossroads and possibly headed toward nationalization. One of the Penn Central trustees, Jervis Langdon, advocates the merger of all of the over 70 large railroads in the United States into a single privately owned entity. He was recently quoted, however, as believing "that the day of the rail merger is probably over" (*New York Times*, March 12, 1971) due to the Penn Central's failure. Despite the failure of the P-C merger and the predictions cited above about the disappearance of rail mergers, the ICC recently approved (December 1971) another large rail combination. Conditional upon certain action, the ICC approved the merger between the Gulf Mobile and Ohio R. R. with the Illinois Central R. R. to form the Illinois Central Gulf R. R.

⁹ See Benham [5], April 5, p. 1.

are exaggerated¹⁰ and not enough attention is given to high labor productivity. Still, anytime a company is saddled with rigid cost structures, in this case due to labor contracts, it is at a distinct competitive disadvantage to other firms and other industries which have less rigid structures. Indeed, recent negotiations and contracts between labor and management in the industry have resulted in changes in the National Work Rules. Management has continually charged that these work rules are extremely damaging to the efficiency of their operations.

□ **Railroad bankruptcy-reorganization duration.** Before we continue our efforts to develop a prediction model for railroads, we should pause to consider a major consequence of declaring bankruptcy under Section 77 of the Bankruptcy Act. An extremely important consideration in a corporation's decision to declare bankruptcy and attempt a reorganization is the *time* it takes to effectively restructure the firm to resume normal operations. In addition, there is always the chance that reorganization efforts will prove fruitless and that a liquidation or a "forced" merger will ensue. In those cases all or most of the reorganization expenses will prove useless, not to mention the delay in receiving the liquidation or merger value of the railroad's assets.

At the time of the enactment of Section 77 in 1933, advocates of the new procedures claimed it would shorten the reorganization period and thus lower its costs. In addition, the old "friendly" receiver would be replaced by an unbiased trustee. Furthermore, it would lead to sounder and more permanent financial structures, thereby lowering the probability of future railroad bankruptcies. While the number of railroad bankruptcies has decreased substantially since 1940, even this fact is illusory. This writer observed approximately "only" 17 significant railroad bankruptcies during the period 1946-1972, ten of which could be considered major (Class I) roads. These figures are misleading, however, if the reader considers that prior railroad liquidations and recent mergers had reduced the number of Class I roads substantially. Also, several of these more recent bankruptcies have occurred in the years 1970-1972, and with continued economic distress many additional roads could be on the path to disaster—an issue to be addressed at a later point.

The first proclamation—that reorganization time would be reduced, and so the consequent costs—has been clearly shown to be invalid. The ICC, in its expressed desire to be perfectly fair and thorough, probably has been one of the prime contributors, along with the federal courts, to lengthy and drawn-out proceedings.

The finding that the Section 77 bankruptcy proceedings are significantly longer than the prior railroad equity receiverships, and also longer than the Chapter X proceedings for manufacturing firms, has significant statistical documentation. Guthmann and Dougall examined 31 major equity receiverships between 1916-1933 and found the average length to be between three and four years.¹¹ They speculated in the 1950s that with the lengthy Section 77 proceedings already observed, the average time would be double the pre-1933 ex-

¹⁰ See Peck [22].

¹¹ In [16], p. 653.

perience (six to eight years). In the present study of railroad bankruptcies we have found the Guthmann and Dougall estimate to be essentially correct. A total of 36 bankruptcies during the period 1938–1970 were traced from their actual petition date to the time the railroad came out of reorganization to resume operations, be liquidated, or be merged. The *mean* bankruptcy-reorganization period was *seven years and seven months* with the *median* period being *seven years and two months*. This is significantly greater than the pre-1933 period and also compares quite unfavorably with an average two years and three months' period for manufacturing companies in bankruptcy.¹²

Admittedly, there should be some tradeoff between the positive effects of reaching a fair and equitable solution and the desire to reduce the time and costs of the bankruptcy process.¹³ In our opinion, the lengthy delays are not necessarily to the benefit of anyone (except perhaps the legal profession and expert witnesses).

Long delays in railroad proceedings are primarily due to: (1) lengthy courtroom hearings and appeals; (2) complexity in railroad capital structures; (3) reluctance of all parties to face the economic situation realistically in the hope for higher earnings; (4) difficulty in arriving at accurate earnings estimates; (5) an overburdened and sometimes overrigid ICC; and (6) the sheer time involved in court review of reorganization plans. These and other reasons were called to the attention of law makers as early as 1940 by Florence Dembitz,¹⁴ and there is no evidence that conditions have improved since then. In conclusion, it is felt that legislation is needed to effect railroad reorganizations on a more rapid basis.¹⁵

■ The statistical technique utilized in subsequent empirical work is called linear discriminant analysis (LDA). While not so popular as a similar multivariate technique, regression analysis, LDA has been utilized in a variety of disciplines since its first application in the 1930s.¹⁶ More recently, this method or a similar multiple variable approach has been applied successfully to financial problems such as consumer credit evaluation,¹⁷ investment classification¹⁸ and, most

3. Discriminant analysis and its potential for railroads

¹² See Altman [1], p. 895.

¹³ The Bureau of Finance of the ICC computed, as of May 1947, that the average allowance for compensation in eleven prior railroad reorganizations was just under \$2 million [Hearings, Committee on Interstate Commerce: H. R. 2298, 80th Congress; 1st Session (1947)]. While we are not aware of any recent statistics, it is fair to assume that average costs have increased.

¹⁴ In [11].

¹⁵ One might advocate some upper time limit on reorganizations; for instance a five year maximum for reorganization proceedings. In the event that the time limit expires, the various security holders would be compensated according to traditional priority patterns based only on the *net liquidation value* of the firm's assets. Obviously, the difficulties in implementing such a limitation and the possible problems due to timing strategies of the various parties involved present serious obstacles to this proposal. We merely submit this notion in the hope that the additional costs involved, which usually exceed several millions of dollars, could be significantly reduced.

¹⁶ See Fisher [14]. For a comprehensive review of early studies using LDA see W. G. Cochran. [9].

¹⁷ The pioneering work utilizing LDA in a financial context was D. D. Durand [12], pp. 105–142.

¹⁸ See Walter [27] and Smith [25].

relevant to this study, in the prediction of future performance of various types of institutions. The latter studies include manufacturing enterprises,¹⁹ Savings and Loan Associations,²⁰ and commercial banks.²¹

LDA is a statistical technique used to classify an observation into one of several *a priori* groupings dependent upon the observation's (e.g., railroad's) individual characteristics. It is used primarily to classify and/or make predictions in problems where the dependent variable appears in qualitative form, e.g., an industrialized vs. undeveloped country, bankrupt or nonbankrupt. Therefore, the first step is to establish explicit group classifications. The number of original groups can be two or more.

After the groups are established, data are collected for the objects in the groups. LDA then attempts to derive a linear combination of these characteristics which "best" discriminates between the groups. If a particular object, for instance a railroad, has characteristics (financial ratios) which can be quantified for all of the companies in the analysis, the LDA determines a set of discriminant coefficients. When these coefficients are applied to the actual ratio, a basis for classification into one of the mutually exclusive groupings exists. The LDA technique has the advantage of considering an entire profile of characteristics common to the relevant firms, as well as the interaction of these properties. A univariate study, on the other hand, can only consider the measurements used for group assignments one at a time.

Another advantage of LDA is the reduction of the analyst's space dimensionality, i.e., from the number of different independent variables to $G - 1$ dimension(s), where G equals the number of original *a priori* groups. This paper is concerned with two groups, consisting of bankrupt railroads on the one hand, and industry averages on the other. Therefore, the analysis is transformed into its simplest form: one dimension. The discriminant function of the form $Z = V_1X_1 + V_2X_2 + \dots + V_NX_N$ transforms individual variable values to a single discriminant score or Z -value which is then used to classify the object, where

$$V_1, V_2, \dots V_N = \text{Discriminant coefficients}$$

$$X_1, X_2, \dots X_N = \text{Independent variables.}$$

The LDA computes the discriminant coefficients, V_j , while the independent variables X_j are the actual values, where

$$j = 1, 2, \dots n.$$

¹⁹ See, Altman [2] where a group of manufacturers ranging up to \$25 million in assets were classified into bankrupt and nonbankrupt groups with 95 percent and 83 percent accuracy one and two statements prior to actual bankruptcy. Later Blum [7] utilized a similar method with legal as well as financial implications. More recently, Robert O. Edmister [13] investigated small firms with the same approach.

²⁰ An early description of the Federal Home Loan Bank Board's Office of Examinations and Supervision efforts in this area is detailed in Leavy [19]. A similar study has been worked out utilizing nonparametric techniques for association ratings in a recent comprehensive work by Ernest Bloch [6]. Significant strides have also been made in the California sector of the Savings and Loan Industry, see [23].

²¹ See Meyer and Pifer [21].

Graphical and statistical presentation of this methodology is described in several standard sources.²²

LDA has the statistical properties and the empirical precedent to make it an appropriate technique for railroad bankruptcy prediction investigation. Before reporting on its railroad application, we will review several of the characteristics of the railroad industry which increase our expectation that an accurate predictive tool can be developed.

□ **Railroad industry advantages.** One particularly fortunate ingredient of a multiple discriminant or even a univariate *railroad* bankruptcy analysis is that we now have a project where the comparative analysis can be concentrated within a relatively homogeneous group of sample firms. As we have pointed out several times, the prior manufacturing study suffered from one potentially serious drawback for broad practical application. The heterogeneous nature of the sample cast suspicion as to its accuracy in predictions of firms within industries where the nature of business operation differs greatly from the average manufacturing firm. The ideal would be to construct individual models for specific industries, or lines of business, so as to remove this drawback. The railroad industry study, therefore, provides an ideal framework from which we can apply a proven technique to an even more appropriate problem area. Specifically, we will match results drawn from a group of bankrupt railroads and compare them with average industry measures drawn from the same time periods as the prior-to-bankruptcy financial ratios. More detailed discussion of this procedure follows shortly. The beauty of this homogeneous type of analysis is that we can be confident that any new firm tested in the model will have similar characteristics that were present in the group of firms used to construct the model. Concentration can be directed specifically to bankruptcy potential without fear of distorted ratio comparisons due to industry differentials. We might add that there are few industries with sufficient bankruptcy experience to provide adequate readily available sample data—a dubious distinction for railroads, but one which can be exploited.

This brings us to another positive reason for our optimism that a viable bankruptcy predictive model in the railroad industry is feasible. Due to ICC reporting requirements, accounting data, both present and past, are for the most part available and quite comprehensive. Although these ICC requirements have changed from time to time, and have been criticized especially for “inflating earnings,”²³ the comprehensiveness of data has remained at a high level and continuity of accounting reporting has only been moderately difficult to interpret. The uniformity of accounting systems required of railroads offers relatively comparable and comprehensive financial and operating data but at the same time puts a burden on the analyst to be selective in his choice of indicators. In addition to the individual firm reporting standards, the ICC provides aggregate annual statistics for

²² See Smith [25], pp. 10–16, and Cooley and Lohnes [10], chapters 6 and 7.

²³ See [24]. We are fully aware of the possibility of misrepresented, mistaken, and manipulative reporting standards. Since we are using secondary data sources, there is a little that can be done about this problem.

the entire railroad carrier industry which is published along with firm data in *Moody's Transportation Manual* on an annual basis.

One of the characteristics of the railroad industry which distinguishes it from most other businesses is its regulatory nature. An appropriate question can be asked as to whether or not a statistical predictive tool can be successfully developed and implemented in such an environment. Substantial evidence from several similar types of industries, at least in terms of the existence of a regulatory governing body, appears to point favorably for the accomplishment of this task. The Savings and Loan industry, for one, has experimented with a multiple discriminant model to be used by their examiners in order to assess the solvency of the member associations around the country. One system, developed at the Federal Home Loan Bank Board, utilized annual and semiannual reported financial data and combined key measures to predict problem areas.²⁴ The entire Federal process is now under review, however, due to data source and timing problems. In the commercial bank sector an accurate predictive model was developed for assessing bank insolvencies by utilizing a type of bivariate regression analysis which is essentially identical to multiple discriminant analysis for a two-group analysis. The model, developed by Professors Meyer and Pifer at the Federal Deposit Insurance Corporation, combines reported financial data ratios with local bank characteristics in order to make solvency predictions.²⁵ In the above instances, the motivation for developing an early warning device was due to the widespread public interest involved as well as to making the economic system as stable as possible. Nothing shatters public confidence in their financial institutions like insolvencies, potentially causing losses of hard earned savings and deposits. The railroad industry and its governing bodies simply cannot sit back and let roads go bankrupt without attempting to provide emergency measures when insolvency becomes a real possibility.

4. Model development

■ A group of twenty-one railroads that went bankrupt between the years 1939–1970 was compiled and scrutinized.²⁶ These railroads include those bankruptcies where sufficient data were available on their prebankruptcy operations. They are comprised of Class I roads (greater than \$5 million in revenues) and/or those which have filed under Section 77 of the Bankruptcy Act. Balance sheet and income statement data were then gathered for one statement and two statements prior to actual bankruptcy. *Moody's* also publishes extremely comprehensive industry data on an annual basis, and the next step was to gather financial statistics for the industry as a whole for the same years that we have bankruptcy data. Industry ratios and measures of performance are derived from aggregated totals for the various balance sheet and income statement accounts. These aggregates include all of the railroads in the industry for a particular year. The objective at this point is to compile a list of ratios for the bankrupt group sample in order to compare average results with that of the

²⁴ See Leavy [19].

²⁵ See [21].

²⁶ See the list in Appendix 1.

include those bankruptcies were sufficient data were available on their aggregate industry data.²⁷

The actual ratios compiled encompass financial measures including those found to be meaningful in prior studies, some popular measures found in the literature, and ratios thought to be particularly relevant to the railroad industry. The ratios are listed in Appendix 2. The measures encompass three overall groupings of financial indicators: (1) liquidity measures; (2) profitability and efficiency measures; and (3) solvency and leverage measures. Several of the ratios are quite similar and any multivariate analysis would probably not include all of these measures.

Our *a priori* feeling is that the liquidity, profitability, and solvency measures of the bankrupt sample will show significantly worse results than the industry averages, and that these differentials will increase as bankruptcy approaches. That is, the two statements prior to bankruptcy measures will be “healthier” than the one statement prior to results for the bankrupt group.

□ **Empirical results.** The first step in our empirical investigation is to examine the average ratios for the bankrupt group of firms and to compare them statistically to the average railroad industry ratios. More precisely, we calculated averages for the 14 ratios listed in Table 1 for the bankrupt group sample of 21 firms based on data drawn from one statement and from two statements prior to bankruptcy. Industry averages were calculated from the same periods. For instance, the sample of bankrupt firms includes three firms who declared bankruptcy in 1970, and therefore data were collected as of the end of 1969 and 1968 for each firm. Likewise, industry average ratios are calculated by giving proportional weighting to the 1969 and 1968 years (three out of 21). Since the bankrupt groups’ petitions covered a relatively long span of time, the proportional weighting helps to remove any bias due to trend movements in the variable values. Ideally, we would prefer bankrupt firm sample points falling within a relatively short period of time especially for a discriminant parametric model.

The ratio results listed in Table 1 conform with our *a priori* expectations and indicate that a multivariate prediction model is a viable possibility. The bankrupt group’s ratios show significantly worse measures (*F*-ratios significant at 0.01 level) than the industry averages (with 3 exceptions) for both one and two statements prior to failure. In addition, the bankrupt averages all show deterioration as failure approaches.

Since most of the ratios show discriminating potential, we will briefly mention just a few of these measures and their characteristics.

²⁷ The objective in assembling a sample of nonbankrupt railroads is to compare their financial measures with that of the bankrupt group. Rather than utilize a group of randomly selected roads for this purpose we chose industry averages for various years in order to represent a “healthy” railroad’s profile. It was our feeling that an individual road’s nonbankruptcy status does not necessarily mean a healthy status. (Admittedly, however, one could argue that the *railroad* industry aggregates are not healthy either.) Also, one randomly selects from a population in order to represent that population. In this case we “select” the population itself. We realize, however, that utilizing industry aggregates to represent healthy railroads adds a potential statistical bias; therefore subsequent tests utilize a randomly selected sample of railroads.

TABLE 1

BANKRUPT FIRMS AND RAILROAD INDUSTRY RATIOS

RATIO NAME AND NUMBER*	ONE STATEMENT PRIOR TO BANKRUPTCY			TWO STATEMENTS PRIOR TO BANKRUPTCY		
	AVERAGE FOR BANKRUPTCY GROUP	AVERAGE FOR INDUSTRY	F RATIO**	AVERAGE FOR BANKRUPTCY GROUP	AVERAGE FOR INDUSTRY	F RATIO**
I LIQUIDITY						
(1) NCA/TA	-11.0%	2.8%	10.4	- 7.3%	2.9%	8.8
(2) NCA/OR	-42.0%	5.7	7.7	-36.6	6.8	6.2
II PROFITABILITY & EFFICIENCY						
(3) EBIT/TA	- 3.0%	5.0%	10.3	- 2.0%	5.0%	9.9
(4) OR/TTP	24.8	31.8	5.7	25.2	31.4	4.1
(5) OR/NTP	30.5	41.5	0.2 ^a	32.0	40.9	0.4 ^a
(6) OPERATING RATIO	87.4	75.4	15.5	85.3	75.1	18.9
(7) TE/OR	44.3	37.2	30.9	45.4	37.1	44.9
(8) EAIT/OR	-15.0	6.4	56.2	-11.0	6.1	75.5
(9) MAIN/TTP	7.3	8.7	0.3 ^a	7.3	8.7	0.5 ^a
(10) GROWTH RATE IN OR	1.6	3.9	0.9 ^a	2.7	3.7	0.2 ^a
III SOLVENCY AND LEVERAGE						
(11) ES/TA	- 31.8%	18.6%	22.4	-27.3%	18.1%	17.3
(12) TD/TA	91.2	50.5	13.8	88.2	50.7	10.9
(13) CHARGES EARNED	- 0.6X	2.3X	54.1	- 0.5X	2.2X	59.4
(14) CASH FLOW/FIXED CHGS.	- 0.5X	2.3X	65.8	- 0.4X	2.2X	74.5
* DUE TO SPACE LIMITATIONS THESE RATIOS APPEAR IN ABBREVIATED FORM AND FOR THEIR FULL DEFINITION PLEASE REFER TO APPENDIX 2						
** THIS RATIO MEASURES THE SIGNIFICANCE OF THE MEAN VALUES BETWEEN THE TWO GROUPS						
^a NOT SIGNIFICANT ON A UNIVARIATE BASIS.						

Both liquidity measures exhibit large bankrupt *vs.* industry differentials, and this writer favors ratio (1) New Current Assets/Total Assets. Only two of the 21 bankrupts showed this ratio as greater than the industry average, while as many as six displayed greater than the industry average for ratio (2). Therefore, the greater volatility of operating revenues, the base in (2), leads us to believe ratio (1) will prove to be a more dependable discriminator. In the profitability sector, ratios (3) and (8) contain only one bankrupt firm which possessed a profitability ratio greater than the industry average at one and two statements prior to bankruptcy. Ratio (3), however, has a far greater *F*-ratio and is more significant on a univariate basis. We also have confidence that an extremely important railroad measure of efficiency—the operating ratio (6)—will provide important discriminating power. This measure, which compares operating expenses with operating revenues, was 12 percent greater in the bankrupt group *vs.* the industry level. Surprisingly, however, four bankrupt firms showed slightly lower than industry average operating ratios which indicates that further analysis on a multivariate basis is required to determine its overall contributory power to bankruptcy prediction.

In the solvency-leverage sectors, all four measures appear to be powerful discriminators. The Earned Surplus/Total Asset ratio (11), which is a measure of cumulative profitability over time, did not con-

tain a single bankrupt firm that even approached the industry average. The Debt-Asset ratio (12) also appears to be extremely powerful, with several of the bankrupt firm ratios rising well over 100 percent. There appears to be little to choose from between the fixed charges earned ratio (13) and the cash flow coverage ratio (14) in their ability to add to a multivariate model. Both appear to be excellent discriminators.

□ **Variable profile analysis.** From the original list of 14 variables, listed in Table 1, the discriminant profile of variables which is selected as providing an accurate railroad bankruptcy prediction model is

$$Z = f(X_{14}, X_7, X_{11}, X_{10}, X_8, X_6, X_3), \quad (1)$$

where:

- X_{14} = cash flow/fixed charges,
- X_7 = transportation expenses/operating revenues,
- X_{11} = earned surplus/total assets,
- X_{10} = 3-year growth rate in operating revenues,
- X_8 = earnings after taxes/operating revenues,
- X_6 = operating expenses/operating revenues, and
- X_3 = earnings before interest and taxes/total assets.

The ordering of these seven variables is derived from the discriminant computer program which selects variables in the order of their contributory importance. The process, a stepwise discriminant analysis,²⁸ first chooses the variable with the greatest F -value (in this case X_{14} with an $F = 65.8$ —see Table 1), and then includes other variables such that the resulting multivariate F -ratio is maximized for that number of degrees of freedom. The multivariate F -ratio is a test to determine the overall discriminating power of the model.²⁹

²⁸ The computer program, BMD07M, is a standard statistical procedure available in most computer centers. Discriminant analysis programs are also available on a nonstepwise basis.

²⁹ This is the common multivariate F -value which is the ratio of the sums of squares between groups to the within groups sums of squares. When this ratio of the form,

$$\lambda = \frac{\sum_{g=1}^G N_g [\bar{y}_g - \bar{y}]^2}{\sum_{g=1}^G \sum_{p=1}^{N_g} [y_{pg} - \bar{y}_g]^2},$$

where

- G = Number of groups,
- g = Group g , $g = 1 \dots G$,
- N_g = Number of firms in group g ,
- y_{pg} = Firm p in group g , $p = 1 \dots N_g$,
- $\bar{y}_g$ = Group mean (centroid),
- $\bar{y}$ = Overall sample mean,

is maximized, it has the effect of spreading the means (centroids) of the G groups apart and, simultaneously, reducing dispersion of the individual points (firm Z -values, y_{pg}) about their respective group means. Logically, this test is appro-

The specific parametric discriminant model selected is

$$Z = 0.2003(X_{14}) - 0.2070(X_7) + 0.0059(X_{11}) - 0.0647(X_{10}) \\ + 0.1040(X_8) + 0.0885(X_6) + 0.0688(X_3). \quad (2)$$

The group means of the two-group railroad sample are

$$(1) \text{ Bankrupt Railroads} = -3.640$$

$$(2) \text{ Industrial Averages} = +0.299,$$

with the F -ratio for this seven-variable profile = 20.1 (significant at the 0.01 level).³⁰ This test rejects the null hypothesis that the observations come from the same population.

□ Variable description.

X_{14} (*cash flow/fixed charges*): This variable is a slight variation of the traditional fixed charges earned ratio (13) and its denomination is in number of times. Both variables X_{13} and X_{14} are extremely significant on a univariate basis (Table 1) with X_{14} slightly more so due to its relatively lower variance within the groups. Because of the two variables' high simple correlation [$r = 0.57$ in Group 1 (Bankrupt) and $r = 0.82$ in Group 2] only one, X_{14} , appears in the final discriminant profile. The cash flow figure is simply the annual depreciation and amortization expense total added to the earnings after taxes. This ratio is also a variation on the recently popularized cash flow-total debt ratio;³¹ X_{14} more closely reflects the railroad's ability to meet its fixed charges.³²

X_7 (*transportation expenses/operating revenues*) and X_6 (*operating ratio*): These measures are particularly unique to the railroad industry as performance indicators. The higher the ratio, the less efficiently the road is usually operating. For the most part, the type of expenses represented in both ratios is variable in nature and falls within the province of operational efficiency. Although the differential between the two groups' average ratio for the operating ratio X_6 is greater than for the transportation expense ratio X_7 —12 percent vs. 7.1 percent—the higher F -ratio for X_7 is due to a considerably smaller standard error—approximately 9 percent for X_6 vs. 4.0 per-

appropriate because one of the objectives of the LDA is to identify and to utilize those variables which best discriminate *between* groups and which are most similar *within* groups.

³⁰ For $F = 7$, 34 degrees of freedom the .01 significance level = 3.21. The F -statistic falls as the number of variables in the profile increases. We observe that the F -ratio = 65.8 with a one variable profile (X_{14}) and diminishes as follows: 50.0, 36.3, 28.1, 23.4, 20.2 until we achieve the 20.1 figure for the entire 7 variable structure.

³¹ Beaver's study [4] concluded that this ratio was the most effective on a univariate basis. Our results are consistent in this case.

³² X_{14} is a type of a margin-of-safety variable, common in fixed income security analysis. Actually the ratio is not a true safety margin measure because the cash flow total already incorporates fixed charges and also is adjusted for taxes which are affected by the tax deductible nature of fixed expenses. Since it was not possible in many cases to adjust the numerator fully, the ratio we used essentially stands as a close proxy for the theoretically correct measure. In an industry where fixed charges play such an important role in the performance of its member firms, it is not surprising to find X_{14} to be such an important indicator of railroad bankruptcy.

cent for X_7 . Discrimination results are slightly less powerful when the model includes only six variables with X_6 excluded.³³

X_{11} (*earned surplus/total assets*): This ratio is another solvency measure which in its purest sense measures the cumulative profitability of the railroad and indicates the book value net worth of the company based solely on past performance. The bankrupt railroad average is -31.8 percent, indicating that those roads prone toward failure have suffered considerable operating losses. This average ratio, however, is not so drastic as the -61.6-percent average for the manufacturing bankrupt group study.³⁴ This can be explained partly by the fact that the average railroad firm in our sample is considerably larger than its manufacturing counterpart and therefore cumulative operating losses, although often large in absolute amounts, are not so significant, relative to total assets. Another possible reason, unfortunately, is the potential mitigating effect on the earned surplus account when a railroad goes through some type of quasi-reorganization, e.g., substitution of senior claims by new equity and cancellation of earned surplus deficits into capital surplus. When this occurs, the earned surplus does not reflect the "true" cumulative earning power.

X_{10} (*three-year compound growth rate in operating revenues*): This flow measure represents an attempt to capture the effects of insufficient revenues, given that most roads have high fixed costs. The debilitating "magnification" of losses effect can result in railroad failures despite a road's long existence and supposedly established reputation. The three-year span is purely arbitrary and is chosen to examine the relatively short-run effects of poor operating performance. Of all the variables included in the variable profile, this one is by far the least significant on a univariate basis, and therefore we do not have confidence in its coefficient (the sign of which is not what one would expect). Yet, we find it does contribute to the overall statistical discriminating power of the entire profile. This characteristic of discriminant analysis—the inclusion of variables which are not univariately significant—has been demonstrated in previous works.³⁵

X_8 and X_3 (*earnings ratios*): Both measures report on the one-year earnings performance of the railroad. While X_8 represents a type of after-tax return on sales (traditionally called a profit margin), X_3 measures the "pure" earning power of the road's assets. Since several railroads diversified in the 1960s in an attempt to protect themselves against economic downturns, these ratios may not

³³ Note that in equation (2) the sign of the coefficient for X_6 is positive and therefore not what one would expect. This is due to the high direct correlation with X_7 which has the expected negative sign. When one combines the two coefficients [$+0.0885(X_6)$ and $-0.2070(X_7)$] the result is very close to the coefficient of X_7 in a variable profile which excludes X_6 . We chose the 7-variable model due to slightly more accurate discrimination results. The 6-variable function is

$$Z = 0.2539(X_{14}) - 0.1442(X_7) + 0.0071(X_{11}) - 0.0632(X_{10}) \\ + 0.0359(X_8) + 0.0434(X_3),$$

with $F = 20.2$.

³⁴ See Altman [2].

³⁵ See Cooley and Lohnes [10], p. 121; Altman [2], p. 597, and Cochran, [9], p. 182.

represent the return on railroad assets alone. This is especially true if the financial statements are on a consolidated basis. Wherever possible, in the original sample and subsequent tests, we used unconsolidated reports—usually those submitted to the ICC.

Variables not included: Two variables are conspicuously absent from the final variable profile. One type of variable missing is a liquidity measure. Neither X_1 nor X_2 managed to enter the profile although both were significant on a univariate basis. This is not terribly distressing since a firm's ability to meet cash requirements is not always reflected in its current position. The second variable not present is the total debt/total asset ratio (X_{12}), although large differences exist between the two groups. Probably this is due to high intragroup dispersion and the fact that similar information is supplied by X_{11} .

□ **Discrimination results.** With the estimation of the discriminant coefficient values, it is possible to calculate discriminant values (Z -scores) for each observation in the sample and to classify them into one of the two initial groups. The essence is to compare the profile of an individual railroad with that of the alternative grouping profiles and to assign it to the group it most closely resembles.

In this type of two-group classification procedure there are two types of errors which are possible. A Type I error is observed when a bankrupt railroad is classified as belonging to the industry or "healthy" group while a Type II error is the reverse: a healthy observation is classified as a bankrupt. Obviously the costs of misclassifying a future bankrupt are considerably greater than the other error type.

□ **One statement prior results.** The empirical classification results of the original two-group sample at one statement prior to bankruptcy are extremely accurate. Only one bankrupt firm was misclassified as healthy while every industrial average observation was correctly classified. This means that the Type I error was just 4.76 percent and the Type II error was zero. The Z -score results for the 42-observation original sample are depicted in Figure 1 and Table 2.

FIGURE 1

RAILROAD DISCRIMINANT SCORES, TWO GROUPS, ONE YEAR PRIOR TO BANKRUPTCY

$$[Z = 0.2003 (X_{14}) - 0.2007 (X_7) + 0.0059 (X_{11}) - 0.0647 (X_{10}) + 0.1040 (X_8) + 0.0885 (X_6) + 0.0688 (X_3)]$$

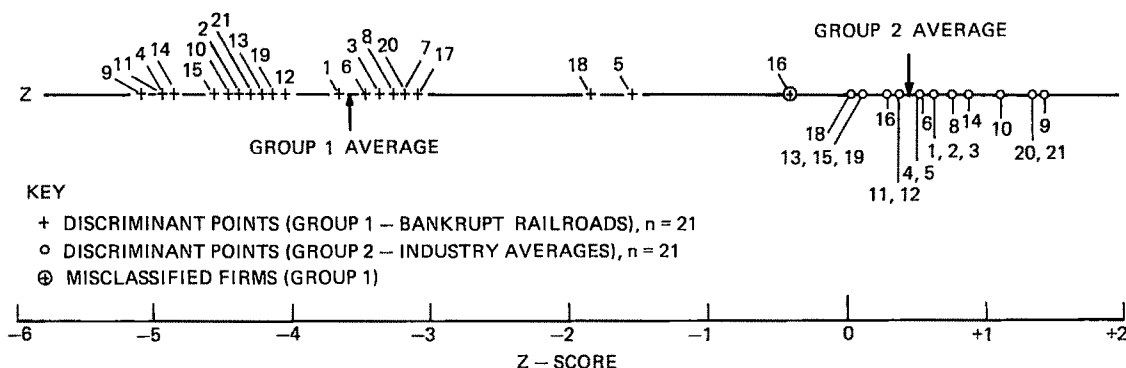


TABLE 2

DISCRIMINANT Z-SCORES — ORIGINAL TWO GROUP SAMPLES

ONE STATEMENT PRIOR TO BANKRUPTCY			TWO STATEMENTS PRIOR TO BANKRUPTCY		
RANK	RR-ID*	Z ₁	RANK	RR-ID*	Z ₂
1	I09	1.4460	1	I13	0.8879
2	I10	1.1805	2	I09	0.8365
3	I14	0.8879	3	I07	0.7246
4	I08	0.7246	4	I06	0.6838
5	I03	0.7021	5	I08	0.6422
6	I02	0.7021	6	I05	0.6371
7	I01	0.7021	7	I04	0.6371
8	I05	0.6985	8	I03	0.6359
9	I04	0.6985	9	I02	0.6359
10	I07	0.6838	10	I01	0.6359
11	I06	0.4821	11	I10	0.6097
12	I12	0.3286	12	I12	0.2759
13	I11	0.3286	13	I11	0.2759
14	I16	0.1573	14	I21	-0.0128
15	I13	0.0312	15	I18	-0.0974
16	I18	-0.1111	16	I17	-0.0974
17	I17	-0.1111	17	I16	-0.1111
18	I15	-0.2374	18	I15	-0.1425
19	I19	-0.2625	19	I14	-0.2249
20	(R16)	-0.3122	20	I20	-0.2783
21	I21	-1.3728	21	(R16)	-0.9703
22	I20	-1.3728	22	I19	-1.3728
23	R05	-1.5587	23	R18	-2.3820
24	R18	-1.7792	24	R08	-2.5546**
25	R17	-3.0904	25	R06	-2.8723**
26	R07	-3.1204	26	R21	-2.9677**
27	R20	-3.2282	27	R02	-3.0652**
28	R08	-3.2978	28	R03	-3.1300**
29	R03	-3.3114	29	R07	-3.2039**
30	R06	-3.4911	30	R20	-3.4050
31	R01	-3.6983	31	R14	-3.4246**
32	R12	-4.0427	32	R19	-3.6909**
33	R19	-4.1445	33	R04	-3.7308**
34	R13	-4.2202	34	R12	-3.9219**
35	R21	-4.3189	35	R01	-3.9537
36	R02	-4.3481	36	R13	-4.0723**
37	R10	-4.3877	37	R17	-4.1210
38	R15	-4.4602	38	R05	-4.3218
39	R14	-4.7429	39	R15	-4.3539**
40	R04	-4.8250	40	R11	-4.5638**
41	R11	-4.8321	41	R10	-4.7693
42	R09	-5.1305	42	R09	-4.8577**

* ID # KEYED TO APPENDIX 1.

** INDICATES HIGHER Z₂ SCORE THAN Z₁.

() INDICATES RAILROAD MISCLASSIFICATION.

In the calculation of Z-scores for each observation, it is important to note that the first variable (X_{14}) is expressed as number of times, e.g. 2.3, while the other six variables are all expressed in "absolute" percent form. For example, X_{10} for a firm may be 35 percent and is expressed as 35.0.

Figure 1, shown in one dimensional space, maps the original sample observations. Note that only one railroad (R-16—Wisconsin Central Railway) which is ranked 20th by its Z-score, falls outside of its group space. We arbitrarily choose a Z-score of -1.465 as the cutoff point between classification into the two groups. This point is simply the midpoint between the lowest industry score (-1.3728 for I 20) and the highest correctly classified bankrupt railroad score

(-1.5587 for R-05). The railroad numbers are keyed to Appendix 1. In total, the discriminant function classified the original sample 97.62 percent accurately (41 out of 42).

Note that our approach for classification is based on a cutoff value. A potentially more powerful and certainly more sophisticated approach would be a likelihood ratio test whereby probability estimates of bankruptcy are given. Unfortunately, the statistical packages available did not have this feature. Furthermore, we have never seen this probabilistic approach used on a practical basis. It is fair to say, however, that anyone following up this study should consider this feature.

□ **Two statement prior results.** The exceptional accuracy of the original sample is matched by the two statement prior to bankruptcy classification. The Z-score and ranking results for this group are also shown in Table 2. Using the original sample coefficients [equation (1)] we take the ratio values of the bankrupt railroads two statements prior and sum the 7 products to arrive at the Z-scores. Note again that with the -1.465 cutoff score, only R-16 is misclassified. A careful comparison between one statement and two statement prior results reveals that in 14 of the 21 bankrupt cases, the firm's Z-score deteriorates as failure approaches (marked by** in Table 2). Still, the fact that one third of the cases (7 of 21) did not display this pattern indicates that the firm will not necessarily deteriorate continuously as bankruptcy approaches.³⁶

□ **An alternative modeling strategy.** Earlier we mentioned reasons for selecting industry aggregate data to compare with that of the bankrupt railroad sample. We also noted the possibility of a statistical bias which may result from this approach. In particular, we may observe significantly different values for the actual ratios, and more than likely the distribution of ratio values within the group will be less dispersed if we utilize industry data for various years than another sampling technique using individual firms representing the nonbankrupt sample. Since tightness of dispersion *within* groups is a positive attribute for good discrimination *between* groups,³⁷ we expect better discrimination than if the dispersion within the nonbankrupt group were more dispersed. This assumed, of course, that the other relevant parameter—mean values—is not significantly different between the two possible nonbankrupt group sampling techniques.

Due to the above considerations, we construct a second model with *individual firms* comprising the nonbankrupt group. For every bankrupt railroad observed in the year of the bankruptcy petition (Appendix 1) we randomly select a nonbankrupt railroad from the nation's Class I roads of that year. The selected firms must have remained solvent in a bankruptcy sense for at least five years after the year they were selected. (This could not be assured, of course, for the most recent years—1967–1970.) For example, two railroads were selected in 1939, one in 1940, . . . , and three roads in 1970. From

³⁶ This is consistent with prior results which showed that bankrupt manufacturers experienced the greatest deterioration in their individual ratios between the third and second statements prior to bankruptcy although most ratio *averages* continued to deteriorate right up to failure. See Altman, [2], p. 605.

³⁷ See note 29.

this new sample of healthy railroads we calculate values for the seven variables discussed above.

The mean values and standard deviations of each variable are displayed in Table 3 for this new sample of roads as well as for the industry aggregates and the bankrupt railroad groups. Note that the

TABLE 3
GROUP MEANS AND STANDARD DEVIATIONS* COMPARISON

VARIABLE	NONBANKRUPT FIRM GROUP	INDUSTRY AGGREGATES GROUP	BANKRUPT GROUP
EBIT/TA (X_3)	5.31 % (3.99)	5.03 (2.0)	-3.02 (11.3)
OPERATING RATIO (X_6)	74.10 % (13.5)	75.43 (5.0)	87.42 (13.0)
TE/OR (X_7)	34.29 % (5.9)	37.22 (3.2)	44.54 (5.1)
EAIT/OR (X_8)	5.40 % (7.9)	6.35 (3.8)	-14.96 (12.5)
GROWTH RATE OR (X_{10})	3.85 % (10.0)	3.92 (7.8)	1.59 (8.3)
ES/TA (X_{11})	25.43 % (25.5)	18.57 (10.9)	-31.82 (37.5)
CASH FLOW/FIXED CHGS (X_{14})	6.88 X (12.1)	2.33 (1.3)	-0.47 (0.9)

*STANDARD DEVIATIONS IN PARENTHESES.

average values for the nonbankrupts are extremely similar to those of the industry aggregates. In fact, at least five of the seven variables have virtually the same mean values. The two variables (X_{11} and X_{14}) which show somewhat different mean values both favor the nonbankrupt firms, but there is no reason to believe this to be indicative. We might find just the opposite results with another sample. The important finding is the similarity in group means.

As expected, we find that in all cases the standard deviation within the industry aggregates sample is less than the comparable figure for the individual firm nonbankrupt group. This is due to the relatively small changes from year to year in the industry figures. We now examine how this difference affects the discriminating power of a model based on bankrupt and nonbankrupt companies.

The Z-score values for the individual firms (both bankrupt and nonbankrupt) and group means are listed in Table 4. Also included are the coefficients used to calculate the Z-scores.

The resulting discrimination power is lower, but only slightly so, than the model utilizing industry data. The F-value for the entire profile equals 13.0 compared to 20.1 for the industry model, but is still significant at the .01 level. Note that the group mean Z-score values are closer together in this new model,

	Z-Score, Nonbankrupt Group Mean	Z-Score, Bankrupt Group Mean
Industry Sample	+0.299	-3.640
Firm Sample	-0.164	-2.695,

also indicating potentially greater overlap between values of firms in the two groups.

Referring to Table 4 and using a cutoff value between the two groups of -1.336 , we still find excellent discrimination. This cutoff

TABLE 4

DISCRIMINANT Z-SCORES — SECOND TWO GROUP SAMPLE*

RANK	RR-ID	Z-SCORE	RANK	RR-ID	Z-SCORE
1	NB-17	1.069	22	R-18	-1.528
2	NB-16	1.053	23	NB-14**	-1.560
3	NB-01	0.765	24	NB-10**	-1.826
4	NB-08	0.677	25	R-07	-2.197
5	NB-05	0.558	26	R-17	-2.372
6	NB-18	0.499	27	R-06	-2.426
7	NB-04	0.326	28	R-03	-2.439
8	NB-12	0.303	29	R-20	-2.474
9	NB-06	0.192	30	R-08	-2.635
10	NB-09	-0.155	31	R-01	-2.847
11	NB-15	-0.200	32	R-10	-2.900
12	NB-11	-0.232	33	R-11	-2.927
13	NB-21	-0.250	34	R-21	-2.949
14	NB-20	-0.451	35	R-12	-3.191
15	NB-03	-0.500	36	R-04	-3.293
16	NB-02	-0.576	37	R-14	-3.300
17	NB-07	-0.645	38	R-02	-3.312
18	R-16**	-0.646	39	R-15	-3.337
19	NB-19	-1.210	40	R-13	-3.374
20	NB-13	-1.236	41	R-11	-3.413
21	R-05	-1.416	42	R-09	-3.623

* $Z = 0.0033 (X_{14}) - 0.1211 (X_7) + 0.0038 (X_{11}) - 0.0510 (X_{10}) + 0.0697 (X_8) + 0.0468 (X_6) + 0.0499 (X_3)$

**MISCLASSIFIED RAILROADS.

NB NONBANKRUPT RAILROADS.

R: BANKRUPT RAILROADS.

value minimizes the number of misclassifications and is the midpoint between the lowest correctly classified nonbankrupt score (-1.236 , NB-13) and the highest correctly classified bankrupt (-1.416 , R-5). The Type I error remains at the same level as in the original model with the same firm (R-16) misclassified. The Type II error rises to two firms (NB-14 and NB-10) vs. no errors in the first model. The total error, therefore, is an impressively low 7.1 percent and the more important Type I error is at a very low level. Note also that the ranking of the Z-values for the individual bankrupt railroads remains quite similar to the first model.

The above results imply that a railroad bankruptcy prediction model consisting of the seven variables noted earlier has important discriminating potential. We now turn to several validating tests based on secondary samples. In all cases we will be testing the *original model*.

□ **Potential bias and validation tests.** The empirical results reported above are indeed impressive, yet we know that they are biased upward for several reasons. The bias is due, in general terms, to the fact that the same observations which are used to construct the model are also the ones classified by the model. One cause of this bias is due to errors in sampling when estimating the means of the two populations upon which the discriminant coefficients are based. The second source of bias is referred to as search bias. This is the act of reducing the original set of variables (14 in our case) to a reduced profile (7)

which is judged to be most effective. This type of “arm-twisting” may result in excellent discrimination results using a subset of variables from the initial sample, but there is no guarantee that it will be effective for other samples or, at the extreme, the total population.

We can conclude from the above that the initial results are biased upward and we attempt to further test the model in order to validate its effectiveness. A method suggested by Frank et al.³⁸ for testing the aforementioned biases is applied to the initial sample. The essence of this test is to estimate parameters using a subset of each of the two original groups and to test the significance of these results (called the *Analysis Sample Test*); then to classify the remainder of the sample based on the parameters established in the analysis sample (called the *Validation Sample Test*). A simple *t*-test is applied to the results to test their significance. This test is of the form:

$$t = \frac{Q - P}{P},$$

where

Q = Proportion of sample observations correctly classified by the discriminant analysis,³⁹

P = Proportion one expects by chance, e.g., assuming the groups are equal in size, as is the case in the railroad samples,
 $P = .5,$

$$P = \sqrt{\frac{P(1 - P)}{n}}, \text{ and}$$

n = Number of observations in the sample.

Five different replications of this test are performed with the analysis sample each time consisting of eleven firms in each group, $n = 22$. These replications are: (1) all odd railroads and their associated industry average; (2) all evens plus # 1; (3) random selection; (4) first eleven observations in each group; and (5) last eleven observations. The validation test sample comprises the remaining ten observations in each replication, $n = 20$. The same variables that appear in equation (1) are present in each of these tests.

The test results, illustrated in Table 5, reject the hypothesis that there is no difference between groups (both analysis and validation tests) and substantiates that the model does possess significant discriminating power on observations other than those used to establish the parameters of the model. Therefore, any search bias does not appear to be significant.

□ **Random selection test.** Thus far, we have shown that the railroad prediction model is extremely accurate in classifying firms in the original sample, both for the entire sample and for various subsets. Another type of test is now performed in order to assess the validity of

³⁸ In [15].

³⁹ A crucial assumption of this test is that Q is generated from random drawings from a binomial population. This is obviously *not* the case in the discrimination results for the sample which determines the model and its coefficients (results reported earlier and the subsequent analysis sample test). The test is valid, however, for the validation test as described.

TABLE 5
VALIDATION SAMPLE TESTS

REPLICATION	ANALYSIS SAMPLE		VALIDATION SAMPLE	
	PERCENT OF CORRECT CLASSIFICATIONS	t-VALUE	PERCENT OF CORRECT CLASSIFICATIONS	t-VALUE
1	100.0	4.67 ^a	85.0	3.13 ^a
2	95.5	4.24 ^a	95.0	4.02 ^a
3	95.5	4.24 ^a	90.0	3.57 ^a
4	100.0	4.67 ^a	80.0	2.68 ^b
5	100.0	4.67 ^a	65.0	1.34 ^c
AVERAGE	98.2		83.0	
OBSERVATIONS	(22)		(20)	
^a SIGNIFICANT AT 0.01.				
^b SIGNIFICANT AT 0.025.				
^c SIGNIFICANT AT 0.25.				

the model on a completely new group of railroads, each randomly selected.

A group of 50 railroads is chosen based on a stratified, random basis. Each sample observation is derived by first randomly selecting a year between 1946–1969. For each year, an observation is then selected from the list of Class I railroads compiled by *Moody's* in that year. For instance, the first railroad selected is the Louisiana & Arkansas Ry. for 1959, denoted as S01. The entire sample is listed in Appendix 3. Finally, data are gathered for each railroad, ratios computed, and Z-scores noted. Note that a firm can, and does, appear more than once in this sample.

In the 50-railroad sample, only six observations had a Z-score below the cutoff point, -1.465 . These include numbers S04, 15, 16, 30, 32, and 39, as denoted in Appendix 3. Of the six railroads classified as potential bankrupts, two (S04 and S16) actually went bankrupt in years subsequent to our test,⁴⁰ one (S39) discontinued all railroad operations, and two others (S30 and S32) were subsequently merged into larger railroad systems; both of which have actually gone bankrupt!⁴¹

One road (S15) is owned by a foreign (Canadian) railway system. In essence, the railroad bankruptcy prediction model, when tested on this new secondary sample, had at most one or two errors of the Type II variety and was correct in predicting bankruptcy

⁴⁰ The New York, New Haven & Hartford R. R. (S04) has a long history of financial problems. It was in reorganization from 1935–1947 and our model predicted bankruptcy in 1960. Some seven months later the company actually petitioned for bankruptcy. It is interesting to note that this same railroad entered our sample another time (S38) and its Z-score in 1950 was $+0.0169$ compared to -3.493 in 1960. The model, therefore, illustrates the New Haven's dramatic demise.

⁴¹ The Boston & Albany R. R. was merged into the New York Central in 1961 and is part of the Penn Central System now in bankruptcy. The Erie R. R. (S32) merged with the Delaware & Lackawanna & Western R. R. in 1960 and became the Erie-Lackawanna Ry. Co. in 1968. Its Z-score in 1957 was -1.676 and the merged road's score in 1970 is -3.329 (see Appendix 4). This road declared bankruptcy on June 24, 1972.

for roads as much as six years (Lehigh Valley R.R.—S16) prior to bankruptcy.⁴²

It is difficult to estimate the Type I error in this randomly selected railroad sample. After a careful investigation of the 50 railroads' histories, we did not find one obvious short term error (that of predicting nonbankruptcy in a near-term failure case). Some interesting patterns do emerge, however, especially with railroads which have recently been cited as potential bankrupts. (These 1970 "potentials" are examined in the next section.) For instance, railroads S17 and S24 had significantly worse Z-scores in 1970 than in 1956 (for the "Milwaukee") and 1964 (for the "Rock Island"). These really cannot be construed as prediction errors since the roads are still solvent and the prediction data were compiled 14 and 6 years prior to 1970, respectively. An extreme case of a "long-term-Type I" error is the Pennsylvania R.R. (S44) which shows a $Z = -1.12$ in 1947; not very "healthy" but still above the cutoff point. Some 23 years later, its merged descendent finally succumbed.

The Maine Central R.R. appears twice, S06 and S28, and its score goes from $Z = +0.840$ in 1954 to $Z = -0.928$ in 1967 despite the discontinuance of passenger operations in 1960. Apparently, this strategy did not improve the situation. In conclusion, it appears that the railroad discriminant model is valid for secondary samples as well as the original one.

□ **1970 class I railroad results and outlook.** As a final test for our railroad bankruptcy prediction model we examined the Class I railroads as of December 31, 1970, and determined that 55 of them could be tested by the discriminant model.⁴³ We then gathered data from their 1970 financial statements and calculated Z-scores for each (listed in Appendix 4). Of the 55 roads tested, 41 had Z-scores *above* -1.465 and were classified as "healthy" railroads. The remaining 14 roads had scores below this level and are listed in Table 6. Five other railroads had scores between -0.3 and -1.465 , and therefore fall within the zone where classification appears to have the highest probability of error.

The 14 roads classified as bankruptcy candidates include six (08, 11, 24, 33, 41, and 43) which are already in bankruptcy. Of the eight remaining low Z-score railroads, at least five (14, 15, 16, 36, and 55) have been designated by railroad analysts as possibly on the brink of failure.⁴⁴ Furthermore, the three remaining failure candidates (03,

⁴² This railroad is 97.33 percent owned by Penn Central and followed its parent into bankruptcy one month later (July 1970). It is interesting to note that its Z-score in 1964, -3.577 , is almost the same as its 1970 score, -3.698 .

⁴³ *Moody's Transportation Manual*, 1970, lists 73 Class I railroads in the U. S. as of December 31, 1969. Three of these roads (Chicago, Burlington & Quincy; Great Northern; and the Northern Pacific) were merged into the Burlington Northern in 1970. The Kansas, Oklahoma & Gulf Railway was merged into the Texas & Pacific Railway in 1970. Hence, we have 70 potential observations. The final sample of 55 is due to (1) lack of data on a few, (2) nonrailway operations of others, and (3) subsidiary roads' not reporting unconsolidated data.

⁴⁴ In the *Barrons* interview [5] the following discussion evolved:

Question—Would you say the Reading (43) is on the brink?

Jansen—It is close to it and has been for years.

Question—How many (others) appear in trouble?

Benham—There are at least half a dozen marginal ones. In the eastern district

TABLE 6
LOW Z-SCORE RAILROADS, 1970

RR#*		Z-SCORE
RR03	ANN ARBOR R R.	-4.384
RR08	BOSTON & MAINE CORP	-5.420
RR11	CENTRAL R R OF NEW JERSEY	-7.490
RR14	CHICAGO & N. WESTERN RY SYSTEM	-1.669
RR15	CHICAGO, MILWAUKEE, ST. PAUL & PACIFIC	-1.930
RR16	CHICAGO, ROCK ISLAND & PACIFIC R R	-3.824
RR24	ERIE-LACKAWANNA RY. CO.	-3.329
RR26	FORT WORTH & DENVER TY	-3.877
RR28	GRAND TRUNK WESTERN R R	-12.950
RR33	LEHIGH VALLEY R R.	-5.841
RR36	MISSOURI-KANSAS-TEXAS R.R	-2.723
RR41	PENN-CENTRAL TRANSPORTATION CO	-5.501
RR43	READING R.R	-2.571
RR55	WESTERN PACIFIC R.R.	-3.467

*KEYED TO APPENDIX 4.

26, and 28) are controlled by larger, more solvent railroads which might explain their nonbankruptcy legal position.

5. Summary and Implications

■ The purpose of this paper has been to discuss the urgent need for an early-warning system covering the historically failure-prone railroad industry and to develop a statistical tool for providing such a system. A multivariate technique called linear discriminant analysis is utilized with published financial data representing the discriminant variables. Empirical results of the original two-group railroad sample are extremely accurate with 97.7 percent correct classification at one and two years prior to bankruptcy. Subsequent replicative and secondary tests provide evidence for the validity of the bankruptcy prediction model.

There are important implications in such a model when used carefully. Internal management control systems would be enhanced by this device, which shows distressing indications at their early stage. Liquidity reserves could be built up (or at least not reduced), and prudent debt policy followed so as not to aggravate the situation. From an external viewpoint, a model of this type could aid state and federal regulatory authorities in their evaluation of individual road situations, leading to prompt action, such as equipment trust certificate endorsement, in order to improve the railroad's profit poten-

you have the Erie-Lackawanna (24). Should that ever go under, it's hard to see how the Delaware & Hudson (19, $Z = -0.612$) can remain unaffected. Moving west, you have the Rock Island (16) which is alive only because of Union Pacific help. There's also the Milwaukee (15), which nobody is helping. You also have the Northwestern (14) as a turn-around situation, but nobody knows how long it can stay turned around. And there's a road like the Katy (36) . . . the only other major situation people are much concerned about is the Western Pacific (55).

The original version of this paper was submitted during the summer 1971, and listed the Erie-Lackawanna (24) and the Reading R. R. (43) as prime bankruptcy candidates. In the paper's subsequent review and revision period both of these roads have in fact petitioned the bankruptcy courts: Reading (1971) and Erie-Lackawanna (1972).

tial. It could also provide a catalyst for intensified merger negotiations or perhaps be used for abandonment petitions. Emergency action on the part of governmental agencies could be instituted early enough to be meaningful, and not construed as a last second, stopgap procedure merely forestalling inevitable failure. The social cost of railroad financial disasters might, in this way, be tempered so as to reduce the devastating effects of major road bankruptcies. Preventive regulation probably carries lower social and economic costs than post mortem reorganization.

Appendix 1

■ ORIGINAL BANKRUPT RAILROAD SAMPLE

<u>Company Name</u>	<u>Bankruptcy Year</u>
1. Lehigh Valley Railroad	1970
2. Penn Central Transportation Company	1970
3. Boston & Maine Corporation	1970
4. Tennessee Central Railroad	1967
5. Central R.R. of New Jersey	1967
6. New York, New Haven & Hartford Railroad	1961
7. Atlantic & Danville Railway	1960
8. Tennessee Railroad	1959
9. New York, Ontario & Western	1954
10. Sacramento Northern Railway	1953
11. Long Island Railroad	1949
12. Huntington & Broad Top Mountain Railroad	1947
13. Missouri & Arkansas Railway	1946
14. Pittsburgh, Shawmut & Northern Railroad	1946
15. St. Johnsbury & Lake Champlain Railroad	1945
16. Wisconsin Central Railway	1943
17. Rutland Railroad	1942
18. Alton Railroad	1942
19. Central of Georgia Railway	1940
20. Florida East Coast Railway	1939
21. Central Railroad of New Jersey	1939

Appendix 2

■ FINANCIAL RATIOS FOR THE RAILROAD BANKRUPTCY STUDY

A. Liquidity Measures

- (1) Net Current Assets/Total Assets
- (2) Net Current Assets/Total Operating Revenues

B. Profitability and Efficiency Measures

- (3) Income Before Interest and Taxes/Total Assets
- (4) Operating Revenue/Total Transportation Property
- (5) Operating Revenue/Net Transportation Property
- (6) Operating Expenses/Operating Revenue (Operating Ratio)
- (7) Transportation Expenses/Operating Revenue

- (8) Income After Taxes and Fixed Charges/Operating Revenue
- (9) Total Maintenance/Total Transportation Property
- (10) 3 Year Compound Growth Rate of Operating Revenue

C. Solvency and Leverage Measures

- (11) Earned Surplus (Balance Sheet)/Total Assets
- (12) Total Debt/Total Assets
- (13) Fixed Charges Earned (Before Taxes)
- (14) Cash Flow/Fixed Charges

Appendix 3

■ RANDOMLY SELECTED RAILROAD Z-SCORES,
1946-1969

<u>R.R. ID NUMBER</u>	<u>RAILROAD</u>	<u>YEAR</u>	<u>Z-SCORE</u>
S01	Louisiana & Arkansas Railway	1959	4.313
S02	Charleston & Western Carolina Railway	1951	1.688
S03	Duluth, Missabe & Iron Range Railway	1952	1.999
S04	New York, New Haven & Hartford Railroad*	1960	-3.493
S05	Spokane Portland & Seattle Railway	1960	-0.309
S06	Maine Central Railroad	1954	0.840
S07	Elgin Joliet & Eastern Railway	1958	2.705
S08	Kansas City Southern Railway	1962	1.545
S09	Norfolk Southern Railway	1949	1.722
S10	Western Railway of Alabama	1961	0.727
S11	Louisiana & Arkansas Railway	1968	3.013
S12	Delaware & Hudson Railroad Corporation	1955	2.941
S13	Chicago & Eastern Illinois Railroad	1962	-0.585
S14	Alabama Great Southern Railroad	1964	3.169
S15	Central Vermont Railway, Incorporated*	1964	-3.618
S16	Lehigh Valley Railroad*	1964	-3.577
S17	Chicago, Milwaukee, St. Paul & Pacific Railroad	1956	0.550
S18	Pittsburgh & Lake Erie	1964	7.638
S19	Green Bay & Western Railroad	1949	10.258
S20	Gulf, Mobile & Ohio Railroad	1963	1.131
S21	Richmond, Fredericksburg & Potomac Railroad	1955	9.676
S22	Louisville & Nashville Railroad	1953	2.386
S23	Elgin, Joliet & Eastern Railway	1952	1.987
S24	Chicago, Rock Island & Pacific Railroad	1964	0.003
S25	Chicago, Burlington & Quincy Railroad	1956	1.468

S26	Central Railroad of Pennsylvania	1951	5.315
S27	St. Louis-San Francisco Railway	1958	0.224
S28	Rutland Railway Corporation	1950	-0.928
S29	Maine Central Railroad	1967	0.553
S30	Boston & Albany Railroad*	1948	-8.473
S31	Chesapeake & Ohio Railway	1965	1.092
S32	Erie Railroad*	1957	-1.676
S33	Florida East Coast Railway	1963	0.618
S34	Western Maryland Railroad	1950	2.302
S35	Wisconsin Central Railroad	1958	-1.149
S36	Long Island Railroad	1963	-1.293
S37	Georgia Southern & Florida Railroad	1950	0.675
S38	New York, New Haven & Hartford Railroad	1950	0.017
S39	New York, Susquehanna & Western Railroad*	1950	-2.231
S40	Savannah & Atlanta Railway	1963	2.741
S41	Illinois Terminal Railroad	1961	0.625
S42	New Orleans & Northeastern Railroad	1966	1.927
S43	Pittsburgh & Lake Erie Railroad	1956	7.225
S44	Pennsylvania Railroad	1947	-1.127
S45	Spokane, Portland & Seattle Railway	1952	-0.032
S46	Norfolk & Western Railway	1965	1.570
S47	Detroit & Toledo Shore Line Railroad	1960	-0.174
S48	Colorado & Southern Railway	1962	0.774
S49	Boston & Maine Railroad	1947	-1.010
S50	Pittsburgh & West Virginia Railway	1958	2.066

* Railroad classified as a potential bankrupt by discriminant model.

Appendix 4

■ Z-SCORES-CLASS I RAILROADS, 1970

RR #	Company Name	Z-Score
RR01	Akron, Canton & Youngstown Railroad	1.785
RR02	Alabama Great Southern Railroad	1.256
RR03	Ann Arbor Railroad	-4.384**
RR04	Atchison, Topeka & Santa Fe Railway	1.887
RR05	Baltimore & Ohio Railroad	-0.896
RR06	Bangor & Aroostock Railroad	2.749
RR07	Bessemer & Lake Erie Railroad	11.935
RR08	Boston & Maine Corporation*	-5.420**
RR09	Burlington Northern Incorporated	-0.262
RR10	Central of Georgia Railway	1.583
RR11	Central Railroad Company of New Jersey*	-7.490**
RR12	Chesapeake & Ohio Railway	0.550
RR13	Chicago & Eastern Illinois Railroad	1.726
RR14	Chicago & North Western Railway System	-1.669**

RR15	Chicago, Milwaukee, St. Paul & Pacific Railroad	-1.930**
RR16	Chicago, Rock Island & Pacific Railroad	-3.824**
RR17	Cincinnati, New Orleans & Texas Pacific Railway	2.593
RR18	Colorado & Southern Railway	0.325
RR19	Delaware & Hudson Railway Company	-0.612
RR20	Denver & Rio Grande Western Railroad	2.527
RR21	Detroit & Toledo Shore Line Railroad	1.380
RR22	Detroit, Toledo & Ironton Railroad	0.415
RR23	Deluth, Missabe & Iron Range Railway	21.588
RR24	Erie-Lackawanna Railway Co.***	-3.329**
RR25	Florida East Coast Railway	2.013
RR26	Fort Worth & Denver Railway	-3.877**
RR27	Georgia Southern & Florida Railway	1.288
RR28	Grand Trunk Western Railroad	-12.950**
RR29	Gulf, Mobile & Ohio Railroad	0.882
RR30	Illinois Central Railroad	0.791
RR31	Illinois Terminal Railroad	-1.432
RR32	Kansas City Southern Railway	0.693
RR33	Lehigh Valley Railroad*	-5.841**
RR34	Louisville & Nashville Railroad	0.457
RR35	Maine Central Railroad	0.004
RR36	Missouri-Kansas-Texas Railroad	-2.723**
RR37	Missouri Pacific Railroad	-0.002
RR38	Monon Railroad	0.035
RR39	Norfolk & Western Railway	0.459
RR40	Norfolk & Southern Railway	1.151
RR41	Penn Central Transportation Company	-5.501**
RR42	Pittsburgh & Lake Erie Railroad	1.256
RR43	Reading Company***	-2.571**
RR44	Richmond, Fredericksburg & Potomac Railroad	4.539
RR45	St. Louis-San Francisco Railway	-0.492
RR46	St. Louis Southwestern Railway	12.245
RR47	Seaboard Coast Line Railroad	0.231
RR48	Soo Line Railroad	-0.241
RR49	Southern Pacific Transportation Company	-0.283
RR50	Southern Railway	1.119
RR51	Texas & Pacific Railway	-0.366
RR52	Toledo, Peoria & Western Railroad	0.933
RR53	Union Pacific Railroad	2.897
RR54	Western Maryland Railway	0.931
RR55	Western Pacific Railroad	-3.467**

* In bankruptcy in 1970.

** Classified as bankrupt.

*** Bankrupt subsequent to 1970.

References

1. ALTMAN, E. I. "Corporate Bankruptcy Potential, Stockholder Returns, and Share Valuation." *Journal of Finance*, Vol. 24, No. 5 (December 1969).
2. ———. "Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy." *Journal of Finance*, Vol. 23, No. 4 (September 1968), pp. 589-609.

3. ———. "Railroad Bankruptcy Propensity." *Journal of Finance*. Vol. 26, No. 2 (May 1971), pp. 333-346.
4. BEAVER, W. "Financial Ratios as Predictors of Failure." *Empirical Research in Accounting, Selected Readings* Chicago: University of Chicago, Institute of Professional Accounting, 1967, pp. 71-111.
5. BENHAM, I. AND JANSEN, A. Interviews in *Barrons* (March 29, 1971 and April 5, 1971).
6. BLOCH, E. "The Settings of Standards of Supervision of Savings and Loan Associations" in Irwin Friend, director, *A Study of the Savings and Loan Industry*, Washington, D. C., July 1969.
7. BLUM, M. "The Failing Company Doctrine." Unpublished Ph.D. dissertation, Columbia University, 1969.
8. BRAY, T. "Bankruptcy: A Chronic Ill for Railroads." *Wall Street Journal* (September 14, 1970), p. 16.
9. COCHRAN, W. G. "On the Performance of the Linear Discriminant Function." *Technometrics*, Vol. 6 (May 1964), pp. 179-190.
10. COOLEY, W. AND LOHNES, P. *Multivariate Procedures for the Behavioral Sciences* New York: John Wiley & Sons, 1962.
11. DEMBITZ, F. "Progress and Delay in Railroad Reorganizations since 1933." *Law and Contemporary Problems* (Summer 1940).
12. DURAND, D. D. *Risk Elements in Consumer Installment Financing* New York: NBER, 1941.
13. EDMISTER, R. O. "An Empirical Test of Financial Ratio Analysis for Small Business Prediction." Paper presented at the Western Finance Association meetings, Vancouver, B. C., August 30, 1971. Published in *Journal of Financial and Quantitative Analysis* (March 1972).
14. FISHER, R. A. "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics*, No. 7 (September 1936).
15. FRANK, R. E., MASSY, W. F. AND MORRISON, G. D. "Bias in Multiple Discriminant Analysis." *Journal of Marketing Research*, Vol. 2 (August 1965).
16. GUTHMANN, H. AND DOUGALL, H. *Corporate Financial Policy* 3d ed. Englewood Cliffs, N.J.: Prentice-Hall, 1953.
17. HILTON, G. W. "The Consistency of the Interstate Commerce Act." *Journal of Law and Economics*, Vol. 9, No. 2 (October 1966), pp. 87-114.
18. "ICC Begins Study of Penn Central Unit, Expects New Reporting, Accounting Rules." *Wall Street Journal* (July 3, 1970), p. 2.
19. LEAVY, H. "Early Warning System—New Supervisory Tool." *Federal Home Loan Bank Board Journal* (September 1969), pp. 12-14.
20. MEYER, J., PECK, M., STENASON, J. AND ZWICK, C. *The Economics of Competition in the Transportation Industries*. Cambridge: Harvard Univ. Press, 1959.
21. MEYER, P. AND PIFER, H. "Prediction of Bank Failures." *Journal of Finance*, Vol. 25, No. 4 (September 1970), pp. 853-868.
22. PECK, M. "Transportation in the American Economy," in S. Harris, ed., *American Economic History*. New York: McGraw-Hill, 1961.
23. PRICE WATERHOUSE & COMPANY. "A Proposed Analytical System to Provide an Objective Basis for Supervision." *State of California Division of Savings and Loan*. December 21, 1967.
24. "Railroads: Numbers Game." *Newsweek* (April 26, 1971), pp. 77-79.
25. SMITH, P. V. *Classification of Investment Securities Using LDA*. Institute Paper #101, Purdue University, Institute for Research in the Behavioral, Economic, and Management Sciences, 1965.
26. VANCE, S. C. "Penn Central—A Lesson for Bank Boards and One-Bank Holding Companies." *The Bankers Magazine* (Winter 1971), p. 78.
27. WALTER, J. E. "A Discriminant Function for Earnings-Price Ratios of Large Industrial Corporations." *Review of Economics and Statistics*, Vol. 41, No. 1 (February 1959), pp. 42-52.
28. WESTON, J. F. "Economic Background of the Penn Central Bankruptcy." *Journal of Finance*, Vol. 26, No. 2 (May 1971), pp. 311-326.

Relative prices on regulated transactions of the natural gas pipelines

Paul W. MacAvoy

Professor of Management
Massachusetts Institute of Technology

and

Roger Noll

Senior Fellow
Brookings Institution

As in many regulated industries, jurisdiction of the commission regulating the natural gas pipelines does not encompass all transactions of the regulated companies. The Federal Power Commission does not regulate direct sales of the pipelines to industry, while it does set profit limits on sales to retail gas utility companies. If regulation is effective, the prices on the regulated sales should be different from those on the unregulated industrial transactions, all else being equal. A model is constructed here of the behavior of the firm with split regulated-unregulated sales. Thenceforth, the model is tested against a sample of paired transactions in the late 1960s. The findings are that price levels on regulated sales would not appear to have been different from those on unregulated sales, after account has been taken of cost and demand differences. The effects of regulation, while in the expected direction, were insubstantial. Moreover, the differences in the institutions of price setting under regulation—in particular, the widespread use of two-part tariffs—if anything, enhanced the ability of the pipelines to charge identical regulated and unregulated prices.

1. Questioning regulatory effectiveness

■ The practices of the Federal Power Commission in regulating gas pipelines can be viewed in two ways. The impression from the Commission's precedent cases has been that of the development of

Paul W. MacAvoy holds the A.B. degree from Bates College (1955) and the M.A. and Ph.D. degrees from Yale University (1956 and 1960, respectively). Professor MacAvoy is Editor of *The Bell Journal of Economics and Management Science*. At the Sloan School of Management, his research centers on the construction of large scale models of energy markets for policy evaluation.

Roger Noll, a Senior Fellow at the Brookings Institution, received the B.S. degree from the California Institute of Technology (1962) and the A.M. and Ph.D. degrees from Harvard University (1965 and 1967, respectively). He is engaged in research on communications policy and the economics of voting patterns in the United States.

This is further work in a series of studies of the Federal Power Commission sponsored by the Brookings Institution under its Program of Studies in the Regulation of Economic Activity. We are grateful to Brookings (and indirectly to the Ford Foundation) for support, and to E. E. Bailey, James Rosse, and David McNicol for counsel. In particular, we wish to express appreciation to Stephen Breyer for his assistance in gaining some rigor and clarity in this member of the series (the book version of which he is a coauthor).

straightforward rules to force prices down to historical average costs. The FPC set the foundations of public utility regulation in this country through its decisions that pipeline profits on original-cost undepreciated capital expenditures should be constrained.¹ This Commission controlled price levels by “formula” allocations of costs plus “fair” profits to specific services and by the requirement that prices not exceed these allocated costs. The famous “Atlantic Seaboard” formula was an outstanding example of stating simple decision rules to limit price regardless of complex (or adverse) effects in particular markets.² But examination of the actual decisions under the rules does not substantiate this impression. From a survey of the cases in the late 1960s, the suspicion is that the rates of return allowed by the FPC did not constrain the profits of the pipeline companies.³ Also, the Commission’s actual rulings in zone and peak pricing cases seem to have had little effect on prevailing price patterns where more than one pipeline sold in a wholesale market—rather, the existence of “another price” in that market predominated over the historical rule.⁴ As a consequence, prices might not have been much reduced by regulation.

The differences between rules and practice are, however, impressionistic; analysis of the structure and level of regulated prices is needed before conclusions can be reached on the results of regulation. The proof lies in the actual behavior of regulated prices—in whether they have been reduced by FPC practice. Fortunately, tests for proof can be done in this case. Whether prices have been affected by the regulatory process can be confirmed or denied by comparing prices subject to FPC regulation with prices on unregulated sales by the same pipeline companies. Such a comparison is made below, beginning with a description of economic models of partial regulation, derived from “A-J-W models” which in turn were partially developed from analyses of regulated pipeline behavior.⁵ The description is for the purpose of finding testable propositions on regulated as compared to unregulated prices. Subsequently, the data on pipeline prices are described and statistical equations are fitted to the data for testing. Evaluation of the equations leads to the conclusion that the results from regulation in the 1960s indeed were quite limited.

■ Markets in which pipelines were sellers in the 1960s can be divided for analytical purposes between those for gas for retail distributors and those for gas as a fuel in industrial furnaces. The first set of markets—the sales to retail public utilities (q_1)—has been regulated by the FPC, while sales to industrial buyers (q_2) have been free of rate-of-return or direct price controls. Regulation is an important factor in pricing. We shall briefly consider the important structural and cost determinants of prices in these markets, as well.

2. Hypotheses on regulated versus unregulated prices

¹ Cf. *El Paso Natural Gas Company* 19 FPC 154 (1958) and *Natural Gas Pipeline Company of America* 40 FPC 91 (1967). More recently and in some detail, cf. *Florida Gas Transmission* 42 FPC 83 (1969).

² Cf. *The Atlantic Seaboard Case* II FPC 43 (1952) and the important discussion of effects in Wellisz [16], p. 33.

³ Cf. MacAvoy [8], pp. 385–387.

⁴ Cf. Wellisz [16]. Cf. also Kahn [6], Vol. 1, pp. 99 *et seq.*

⁵ After Averch and Johnson [1] and Wellisz [16], p. 33.

The price of a service, and consequently the quantity demanded, depend on the extent of pipeline control of the market. The pipelines could control prices on both sets of sales where there were "few" sources of supply, if these "few" shared markets or allocated contract deliveries among themselves. Such coordination might not have been achieved, however, because the allocation scheme derived from voluntary agreement among a number of individual pipelines was not "workable." That is, price control does not work because of the absence of joint price-setting mechanisms or because too many firms are involved to enforce the prices set by means of a "good" decision-making mechanism. Furthermore, even a sound allocation scheme might not be effective, particularly on industrial sales, because of extensive interfuel price competition for this group of consumers. The important determinants on the demand side, however strong their effect, include number of firms, the nature of the price-setting mechanism, and the closeness of substitute forms of energy.

The levels of prices depend also on production and transmission conditions that fix costs on the supply side. The pipeline enterprise incurs total costs C for operating expenses to provide $q_1 + q_2$, and to amortize the capital costs of existing capacity K . The amounts of these costs vary with the distance, and the volume of the gas transported, in large part. Thus, there are three types of constraints on the pipeline's control of prices: the first, elements of FPC price control; the second, those implied by the extent of alternative sources of supply and of mechanisms to control supply; and third, those inherent in cost conditions.

□ **Prices without regulation.** The first model is based on markets with collusive prices but no regulatory control, so as to show the effects of adding both interpipeline rivalry and these controls. Cooperation establishing price control among pipeline sellers could be the rule, so that prices are "monopolistic." Differences in demand conditions and the absence of "leakage" of sales from the low-priced buyer to the high-priced buyer⁶ would make discrimination—charging different prices for different volumes sold to the same and to different consumers—the most profitable policy. The most straightforward means would be to set two-part prices, with α_1 demand charge and P_1 commodity charge for the typical retail public utility buyer, and α_2 demand charge along with P_2 commodity charge for the typical industrial buyer. This would make the seller's profits equal to

$$R = [\alpha_1 + P_1 q_1] + [\alpha_2 + P_2 q_2] - C(q_1, q_2, K), \quad (1)$$

subject to the capacity constraint

$$(q_1 + q_2) \leq K. \quad (2)$$

To maximize these profits, relative commodity changes should be⁷

$$P_1 = P_2 + [mct(q_1) - mct(q_2)], \quad (3)$$

⁶ Leakage from one buyer to another would be impossible without a pipeline; if it occurred, then the "reseller" would be a pipeline chiseling on the agreed structure of prices. This is ruled out here by assumption.

⁷ This follows closely on Baumol and Klevorick [4] and Pressman [13]. Here the discussion is centered on pricing, rather than input factor ratios, for testing purposes, so that propositions appear to be different while they are not.

where $mct(q_1)$ and $mct(q_2)$ are marginal costs of transporting specific amounts, q_1 and q_2 .⁸ That is, commodity prices on volumes to the retail distributor and industrial buyer should differ only by differences in the respective marginal costs of transport.

Pricing patterns differ when there is not this degree of control of markets. The effects of lack of cooperation among pipelines could take many forms, but among the most likely would be the disappearance of such extensive price discrimination as that by two-part tariff. This would be most likely—the buyer from one pipeline paying a high initial charge α would be lured away once and for all by another pipeline with a lower initial charge—because there would be no effective counter-offer from the first pipeline (as there would be to reduced commodity charges). In fact, the elaborate structure of individual demand and commodity charges would likely reduce to single commodity prices, so that the pipeline would operate as if limited to revenues

$$R^* = P_1 q_1 + P_2 q_2 - C(q_1, q_2, K), \quad (4)$$

where P_1 and P_2 are each commodity prices with no demand prices, and revenues are subject to the capacity constraint

$$q_1 + q_2 \leq K. \quad (5)$$

The maximum profit prices on the two classes of service would be such that

$$P_1(1 + 1/e_1) = P_2(1 + 1/e_2) + [mct(q_1) - mct(q_2)], \quad (6)$$

where e_1 and e_2 denote the price elasticities of demand in the two markets.⁹ Thus, the relative prices would be

$$P_1 = \frac{(1 + 1/e_2)}{(1 + 1/e_1)} P_2 + \frac{1}{(1 + 1/e_1)} [mct(q_1) - mct(q_2)]^{10} \quad (7)$$

⁸ The first order conditions for maximum profits R , subject to the constraint $(q_1 + q_2) \leq K$, include the price difference $P_1 - P_2 = \partial C / \partial q_1 - \partial C / \partial q_2$, the difference in costs of transport at the margin. Under conditions in which the firms can discriminate on prices for different amounts consumed by the buyer, the change in revenues on intramarginal units equals $q_1 \partial P_1 / \partial q_1 = \partial \alpha_1 / \partial q_1$, the change in demand charges. That is, what is lost from reductions in price on intramarginal sales can be made up in increases in the initial charge. Thus, only the commodity prices enter into comparisons across classes of service.

⁹ That is, first order conditions for maximum R^* subject to $q_1 + q_2 \leq K$ include $P_1(1 + 1/e_1) - P_2(1 + 1/e_2) = \partial C / \partial q_1 - \partial C / \partial q_2$, where $\partial C / \partial q_1$ and $\partial C / \partial q_2$ are marginal costs of transport for the two classes of service.

¹⁰ An alternative formulation of the same condition as appears in (7) is as follows:

$$MR_1(q_1) - mct(q_1) = MR_2(q_2) - mct(q_2)$$

where $MR_1(q_1)$ and $MR_2(q_2)$ are the marginal revenues of the two services at outputs q_1 and q_2 . (The marginal revenue is the effect on total revenue if the firm attempts to sell one more unit of output. Since it takes account of the fact that the firm will have to lower its price to achieve greater sales, the marginal revenue is less than the price.) This equation requires that the difference between marginal revenue and marginal costs be the same for both services, and equal the value of a unit of capacity.

If all capacity is used in producing the output that maximized profits and if a larger capacity than K could produce even larger profits, then the first order conditions also require that marginal revenues exceed marginal costs for each service. Otherwise, marginal revenues must equal marginal costs in both markets.

and the total quantities of gas in the two categories would be

$$q_1' + q_2' \leq K'. \quad (8)$$

The ability of the sellers to classify buyers as "retail utility" and "industrial" results in price differences that depend on differences in the marginal costs of transportation, and on relative elasticities of demand. This is a level of ability below that of being able to set both demand and commodity prices on each class, and the result is a departure of prices from strict cost differences. There are further departures introduced by regulation.

□ **Prices when sales to retail utilities are regulated.** The effects of regulation depend on the nature of the particular controls applied by the regulatory commission. The FPC sets limits on profits through the artifice of finding fair rates of return on the rate base or investment. We will simplify that construct here by assuming that the Commission directly limits total profits on regulated sales q_1 to M , so that $p_1 q_1 - \gamma C \leq M$, where γ is the proportion of total costs attributed to regulated sales.¹¹ The profits on industrial sales are not regulated at all, so that total revenues are subject only to this partial regulatory constraint. The conditions for maximum profits subject to this and the capacity constraint specify the two prices as¹²

$$P_1 = \delta \frac{(1 + 1/e_2)}{(1 + 1/e_1)} P_2 + (\delta - \gamma\delta + \gamma) \frac{1}{(1 + 1/e_1)} [mct(q_1) - mct(q_2)] \quad (9)$$

¹¹ In the general case, neither γ nor M is fixed. Regulators normally permit firms to earn a rate of return, r , on fixed capital investment, K , so that profits may not exceed rK . Since firms are free to pick K , the limit on profits is variable. In the pipeline industry, capacity has been more than ample to meet demand since the last wave of capital investment ended in the early 1960s, so that K and, hence, M have been fixed.

The assumption that γ is constant is more an assumption of convenience than a description of reality. According to FPC policy, as enunciated, for example, in the Atlantic Seaboard formula, γ depends upon the fraction of total output that is accounted for by regulated sales. A more realistic model would set $\gamma = q_1/(q_1 + q_2)$. In practice, γ has varied very little during the period under examination here, so that little is lost in descriptive content (but much is avoided in complications of estimation) by assuming γ to be fixed.

¹² The reasoning is as follows: maximizing revenues subject to the two constraints calls for maximizing

$$L = P_1 q_1 + P_2 q_2 - C(q_1, q_2, K) + \lambda [M - P_1 q_1 + \gamma C(q_1, q_2, K)] + \mu (K - q_1 - q_2),$$

where λ and μ are regulatory and capacity constraints respectively. The Kuhn-Tucker conditions for a maximum include:

$$\begin{aligned} (1 - \lambda)P_1(1 + 1/e_1) - (1 - \lambda\gamma)\partial C/\partial q_1 &= \mu \quad \text{or} \quad q_1 = 0 \\ P_2(1 + 1/e_2) - (1 - \lambda\gamma)\partial C/\partial q_2 &= \mu \quad \text{or} \quad q_2 = 0 \\ q_1 + q_2 &= K \quad \text{or} \quad \mu = 0 \\ P_1 q_1 - \gamma C &= M \quad \text{or} \quad \lambda = 0. \end{aligned}$$

Assuming that both services are profitable at some level of output and that both constraints are binding (i.e., capacity is not superfluous and the profits allowed by regulation fall below the maximum possible unregulated profits), then these

and the quantities as

$$q_1'' + q_2'' = K'' \quad (10)$$

The equation for relative prices under regulation can be compared with that for the same market and production conditions but without regulation. They differ by the factor δ in the coefficient of P_2 and by the factor $(\delta - \gamma\delta + \gamma)$ in the coefficient of the marginal cost difference $[mct(q_1) - mct(q_2)]$.¹³ The δ factor is the "regulatory con-

conditions imply:

$$P_1 = \frac{1}{(1-\lambda)} \frac{(1+1/e_2)}{(1+1/e_1)} P_2 + \frac{(1-\gamma\lambda)}{(1-\lambda)} \frac{1}{(1+1/e_1)} (\partial C/\partial q_1 - \partial C/\partial q_2)$$

and

$$q_1 + q_2 = K$$

$$P_1 q_1 - \gamma C = M.$$

In the text, $\delta = 1/(1-\lambda)$, for simplicity of interpretation of the statistical results given below.

The equation for P_1 is somewhat different if γ is allowed to vary with relative output, i.e., $\gamma = q_1/(q_1 + q_2)$. The expression that a firm seeks to maximize is then:

$$L = P_1 q_1 + P_2 q_2 - C(q_1, q_2, K)$$

$$+ \lambda \left[M - P_1 q_1 + \frac{q_1}{q_1 + q_2} C(q_1, q_2, K) \right] + \mu (K - q_1 - q_2).$$

The Kuhn-Tucker conditions for a maximum are:

$$(1-\lambda)P_1 \left(1 + \frac{1}{e_1}\right) - \left(1 - \lambda \frac{q_1}{q_1 + q_2}\right) \frac{\partial C}{\partial q_1}$$

$$+ \lambda \left(1 - \frac{q_1}{q_1 + q_2}\right) \frac{C(q_1, q_2, K)}{q_1 + q_2} = 0 \quad \text{or} \quad q_1 = 0,$$

$$P_2 \left(1 + \frac{1}{e_2}\right) - \left(1 - \lambda \frac{q_1}{q_1 + q_2}\right) \frac{\partial C}{\partial q_2}$$

$$- \left(\frac{\lambda q_1}{q_1 + q_2}\right) \left[\frac{C(q_1, q_2, K)}{q_1 + q_2} \right] = 0 \quad \text{or} \quad q_2 = 0,$$

and

$$M - P_1 q_1 - \frac{q_1}{q_1 + q_2} C(q_1, q_2, K) = 0 \quad \text{or} \quad \lambda = 0,$$

$$K - q_1 - q_2 = 0 \quad \text{or} \quad \mu = 0.$$

Assuming that both q_1 and q_2 are profitably produced, then combining the first two conditions yields:

$$P_1 = \frac{1}{1-\lambda} P_2 \frac{\left(1 + \frac{1}{e_2}\right)}{\left(1 + \frac{1}{e_1}\right)} + \left(\frac{1}{1-\lambda}\right) \frac{\left[1 - \lambda \left(\frac{q_1}{q_1 + q_2}\right)\right]}{\left(1 + \frac{1}{e_1}\right)}$$

$$\times [mct(q_1) - mct(q_2)] - \left[\frac{\lambda}{(1-\lambda) \left(1 + \frac{1}{e_1}\right)}\right] \left[\frac{C(q_1, q_2, K)}{q_1 + q_2}\right].$$

This equation is the same as used in the text, except P_1 now depends on average cost, in addition to P_2 , $mct(q_1)$, and $mct(q_2)$.

¹³ In expression (9), δ and $(\delta - \gamma\delta + \gamma)$ represent "regulatory constraint" effects on prices and outputs. The values they can take can be inferred from the first and second order conditions for maximization of L . As an alternative to (7), the first order conditions can be written:

$$MR_1(q_1) - mct(q_1) - \lambda [MR_1(q_1) - \gamma mct(q_1)] - \mu = 0$$

$$MR_2(q_2) - mct(q_2) + \lambda mct(q_2) - \mu = 0,$$

straint factor" that shows the increase in relative regulated price resulting from the limit on allowed profits from regulated sales.¹⁴ This δ factor is positive and greater than 1.0, and is larger the more stringent the regulatory constraint. With strong regulatory control of

where $MR_x(q_x)$ is the marginal revenue of service X at output q_x . Then, the second order conditions become:

$$\frac{dMR_1(q_1)}{dq_1} - \frac{\partial mct(q_1)}{\partial q_1} - \lambda \left[\frac{dMR_1(q_1)}{dq_1} - \gamma \frac{\partial mct(q_1)}{\partial q_1} \right] \leq 0$$

$$\frac{dMR_2(q_2)}{dq_2} - \frac{\partial mct(q_2)}{\partial q_2} + \lambda \gamma \frac{\partial mct(q_2)}{\partial q_2} \leq 0.$$

If marginal transport costs for the regulated service are constant as output changes, then the relevant second-order constraint reduces to:

$$(1 - \lambda) \frac{dMR_1(q_1)}{dq_1} \leq 0.$$

Since the marginal revenue curve is always negatively sloped, this requires that $\lambda \leq 1$ and, since of necessity $\lambda \geq 0$, then $\delta \geq 1$.

If marginal costs of the regulated service are falling as output increases, the first of the second-order conditions can be written:

$$(1 - \lambda) \left| \frac{dMR_1(q_1)}{dq_1} \right| \geq (1 - \lambda \gamma) \left| \frac{\partial mct(q_1)}{\partial q_1} \right|.$$

If $\left| \frac{dMR_1(q_1)}{dq_1} \right| < \left| \frac{\partial mct(q_1)}{\partial q_1} \right|$, then, for the above to be satisfied, $1 - \lambda > 1 - \lambda \gamma$. But this would require that $\lambda(\gamma - 1) > 0$ and, since $\gamma < 1$ and $\lambda \geq 0$, this is not possible; hence

$$\left| \frac{dMR_1(q_1)}{dq_1} \right| \geq \left| \frac{\partial mct(q_1)}{\partial q_1} \right|.$$

In this case, rearranging the second-order condition,

$$\lambda \leq \left| \frac{dMR_1(q_1)}{dq_1} \right| - \left| \frac{\partial mct(q_1)}{\partial q_1} \right| \bigg/ \left| \frac{dMR_1(q_1)}{dq_1} \right| - \gamma \left| \frac{\partial mct(q_1)}{\partial q_1} \right|$$

Since $\gamma < 1$, the denominator of the above must exceed the numerator, so this requirement implies $\lambda \leq 1$ and thus $\delta \geq 1$.

If marginal costs rise as output increases, λ can exceed unity. The two constraints require only that

$$\lambda \leq \left| \frac{dMR_2(q_2)}{dq_2} \right| + \left| \frac{\partial mct(q_2)}{\partial q_2} \right| \bigg/ \gamma \left| \frac{\partial mct(q_2)}{\partial q_2} \right|$$

and that

$$\lambda \leq \left| \frac{dMR_1(q_1)}{dq_1} \right| + \left| \frac{\partial mct(q_1)}{\partial q_1} \right| \bigg/ \left| \frac{dMR_1(q_1)}{dq_1} \right| + \gamma \left| \frac{\partial mct(q_1)}{\partial q_1} \right|.$$

Since $\gamma < 1$, the right side of both inequalities is greater than unity, so that $\lambda > 1$ and $\delta < 0$ is possible.

¹⁴ From the preceding footnote, $\delta = 1/(1 - \lambda)$. The λ constraint, equal to $\partial L/\partial M$, takes values between 1.0 for stringent regulation and 0.0 for no regulation when the company operates in the "elastic" range of the demand function for sales to public utilities. Thus, $1/(1 - \lambda) > 0$ or $\delta > 0$ when $e_1 < -1$. Under more severe regulatory constraint on prices directly, the company could experience λ greater than 1.0 and operate in the inelastic range of demands for regulated service; then $1/(1 - \lambda) < 0$ and $\delta < 0$ when $e_1 > -1$. In both cases, the coefficient of P_2 , $\delta \frac{(1 + 1/e_1)}{(1 + 1/e_2)}$, is positive. This can be shown as follows. First, $(\delta - \gamma\delta + \gamma)$ must have the same sign as δ . This can be seen by rewriting the first inequality constraint on λ as follows:

$$\gamma \leq \frac{1}{\lambda} + \frac{1}{\lambda} \left| \frac{dMR_2(q_2)}{dq_2} \right| \bigg/ \left| \frac{\partial mct(q_2)}{\partial q_2} \right|,$$

profits, the $(\delta - \gamma\delta + \gamma)$ coefficient also becomes quite large, so that transport cost differences are magnified in regulated-unregulated price differences.

The effect of imposing regulation can easily be to raise the *relative price* for the regulated service. Suppose that in the absence of regulation, the profit-maximizing strategy for the firm is to produce output $q_1 + q_2$ that fully utilizes capacity K . Then, when regulation is imposed, the firm is required to reduce the profits from its regulated service. Since the old prices, p_1 and p_2 , and outputs, q_1 and q_2 , were optimal, profits on the regulated service can be reduced through either increases or decreases in sales.

If the firm responds to the imposition of regulation by lowering prices and increasing sales in the regulated market, then, because of the capacity constraint, sales must be reduced and prices increased in the unregulated market. But, since profits were maximal before regulation was imposed, such a change would cause profits to fall in the unregulated market as well. Consequently, the firm will have

so that, since $\lambda \geq 0$, $\gamma \geq \frac{1}{\lambda}$. Substituting $\lambda = 1 - \frac{1}{\delta}$ this becomes $\gamma \leq \frac{1}{1 - \frac{1}{\delta}}$.

Rearranging, this is $1 \geq \gamma - \frac{\gamma}{\delta}$. If $\delta < 0$, this becomes $\delta - \delta\gamma + \gamma \leq 0$. If $\delta > 0$, this becomes $\delta - \delta\gamma + \gamma \geq 0$. Second, if $\delta < 0$, the demand for the regulated service must be inelastic. The first order condition requires:

$$MR_1(q_1) = \frac{1}{1 - \lambda} mct(q_1)(1 - \lambda\gamma) + \frac{\mu}{1 - \lambda}.$$

If $\delta < 0$, these terms must have the following signs: $\frac{1}{1 - \lambda} < 0$; $mct(q_1) > 0$;

$1 - \lambda\gamma > 0$ (from above when $\gamma \leq \frac{1}{\lambda}$); $\frac{\mu}{1 - \lambda} < 0$ (since $\mu \geq 0$ and $\lambda > 1$).

Hence, $MR_1(q_1) < 0$, which means that the demand is inelastic, or $0 \geq e_1 > -1$,

(or $(1 + \frac{1}{e_1}) < 0$ in (9)). Thus, the coefficient on P_2 has two negative terms if

$\delta < 0$. Note that, by similar argument involving the first order condition on q_2 , $0 \geq e_2 > -1$ in all cases. Hence, the coefficient on P_2 is positive. Similarly, since $\delta - \gamma\delta + \gamma < 0$ if $\delta < 0$ the coefficient on $[mct(q_1) - mct(q_2)]$ must also be positive, since it also has two negative terms.

Finally, if $\delta > 0$ (and $\lambda < 1$), the demand for the regulated service must be elastic (i.e., $e_1 < -1$). Again using the first-order condition $\frac{\partial L}{\partial q_2} = 0$, all of the terms, including $MR_1(q_1)$, must be nonnegative. Thus, $(1 + \frac{1}{e_1}) > 0$ if $\delta > 0$, and all terms in the coefficient of P_2 are positive. And, if $\delta - \gamma\delta + \gamma < 0$, then $\delta < \frac{-\gamma}{1 - \gamma}$, which is inconsistent with $\delta > 0$. Thus, the coefficient on $[mct(q_1) - mct(q_2)]$ also must be positive.

The same reasoning can be applied to the model developed in footnote 12 in which γ was allowed to depend upon relative output. By similar argument, the coefficient on average cost in the equation for P_1 derived from that model must be nonpositive, i.e.,

$$-\frac{\lambda}{1 - \lambda} \left(\frac{1}{1 + \frac{1}{e_1}} \right) \leq 0.$$

This means that the lower P_1 is, relative to P_2 , the greater is average cost, given all the other elements of the equation (including marginal cost).

greater total profits if it leaves the service in the unregulated market unchanged while raising prices and lowering output in the regulated market. Only if the profit limit cannot be obtained by price increases will the firm choose to lower prices for the regulated service.

If capacity was optimal before regulation, and if the regulated price increases, then regulation will cause excess capacity that can be used for more unregulated sales. If capacity was less than optimal before regulation, the regulated firm will, if it increases prices and reduces output for regulated service when regulation is imposed, also lower prices and increase sales in the unregulated market. This is because, in the preregulation situation, if capacity is less than the amount that would yield highest profits, marginal revenues will exceed marginal costs in both markets (see footnote 10). Then, if new capacity becomes available, as it would for the unregulated service if the regulated service were cut back, it is profitable to expand unregulated sales through a reduction in price.

In the long run, there would be a number of impelling reasons to increase capacity. Urgings of the Commission and opportunities to increase total profits M (as a result of higher regulated "rate base" costs) would result in greater capacity than without any regulation.¹⁵ More of the additional capacity would go into unregulated sales because marginal output in the unregulated sector would now be relatively more profitable. Both regulated and unregulated prices would be reduced as a result of capacity expansion, but the regulated price would be relatively higher since these sales would increase relatively less. In other words, the regulated sector with reduced profits

¹⁵ Merely setting limits on total profits M can be compared with regulation by setting a maximum rate of return on capital. This rate-of-return procedure has the effect of enhancing the relative differences between unregulated and regulated prices shown above, at least when the procedure results in returns greater than the economic costs of capital. Consider the profit function $R' = P_1q_1 + P_2q_2 - r_1x_1 - r_2x_2$, where x_1 are capital resources and x_2 noncapital resources available to the regulated firm at prices r_1 and r_2 respectively. The restraint on profits is that $P_1q_1 - \gamma s_1x_1 - \gamma r_2x_2 = M$, where s_1 is the allowed rate of return on capital such that $s_1 - r_1 = V \geq 0$. The first order conditions for maximizing

$$R = P_1q_1 - P_2q_2 - r_1x_1 - r_2x_2 - \lambda(q_1 + q_2 - x_1) - \mu(P_1q_1 - \gamma s_1x_1 - \gamma r_2x_2 - M)$$

include:

$$(1 - \lambda)P_1(1 + 1/e_1) - (1 - \gamma\lambda)(r_1\partial x_1/\partial q_1 + r_2\partial x_2/\partial q_1) - (1 - \gamma\lambda)(s - r_1)\partial x_1/\partial q_1 = 0$$

and

$$P_2(1 + 1/e_2) - (1 - \gamma\lambda)(r_1\partial x_2/\partial q_2 + r_2\partial x_2/\partial q_2) - (1 - \gamma\lambda)(s - r_1)\partial x_1/\partial q_2 = 0.$$

If marginal costs $\partial C/\partial q_i$ are defined as $(r_1\partial K/\partial q_1 + r_2\partial L/\partial q_2)$, then

$$P_1 = \frac{1}{(1 - \lambda)} \frac{(1 + 1/e_2)}{(1 + 1/e_1)} P_2 + \frac{(1 - \gamma\lambda)}{(1 - \lambda)} \frac{1}{(1 + 1/e_1)} \left\{ \frac{\partial C}{\partial q_1} - \frac{\partial C}{\partial q_2} \right\} + \frac{(1 - \gamma\lambda)}{(1 - \lambda)} \frac{1}{(1 + 1/e_1)} \cdot (s - r_1) \cdot \left\{ \frac{\partial C}{\partial q_1} - \frac{\partial C}{\partial q_2} \right\},$$

the same expression for relative prices as in the text, but with the addition of the last term on the right-hand side of the equation. This last term is greater than zero if the marginal cost difference is greater than zero, so that the relative price difference is enhanced when the rate-of-return limits are set on regulated sales. It is as if there were a premium on costs for the excess allowed return; thus, whatever the effects on relative regulated prices from the regulatory constraint, these effects are greater under *this* regimen.

should experience a long-run relative decline in sales and relative increase in price.

□ **Pricing summary.** At this point, the array of theoretical price variations might well be summarized.¹⁶ The price hypotheses differ from each other when there are differences assumed in the degree of market control and in the extent of regulation. When it is possible for pipelines to charge a different price for each unit of gas sold to each retail distributor or industrial buyer, and there is no regulation, the coefficients a and b in the equation

$$P_1 = aP_2 + b[mct(q_1) - mct(q_2)] \quad (11)$$

take on values of 1.0 where P_2 is the price on a comparable industrial sale and $mct(q_1)$ and $mct(q_2)$ are pipeline transport costs to the retail utility and industrial buyers, respectively. When, for reasons of rivalry between pipelines in the same market, prices for all amounts of gas turn out to be uniform to all industrial or all retail utility buyers, but still can be set separately for the two classes, the elasticities of demand e_1 and e_2 enter coefficients a and b so that

$$a = (1 + 1/e_2)/(1 + 1/e_1) \quad \text{and} \quad b = 1/(1 + 1/e_1).$$

There is contrast created by the effects of regulation as well. Regulation results in testable differences in coefficients—the “regulatory constraint” terms δ and γ are found in a and b . The δ and γ terms normally can be expected to increase the regulated price relative to the unregulated price, unless firms experience scale diseconomies or face regulation so severe that they must operate on the inelastic portion of their demand curves.

¹⁶ So far, conditions for relative prices have been derived assuming that demands and costs are known and that regulation is continuous as well as certain. None of these conditions holds, and models with uncertainty should be constructed, if they add greatly to the predictability of the certainty models. But, in fact, at this point, of the many promising uncertainty analyses, those based on monopoly market conditions with price as the decision variable differ little from the certainty-static models. (Cf. the incisive review article by McCall [10], in which he notes that “the strong sensitivity of results to stochastic assumptions is a clear warning that model building in probabilistic microeconomics requires extreme caution and good sense,” at 418.) With these analyses adapted to gas pipeline conditions—with stochastic variation in demand in all markets but greater variation where gas demand is highly elastic with respect to prices of other sources of energy—the resulting prices under uncertainty would be lower than with certainty. (Cf. Mills [11] and Leland [7]). These conditions occur most frequently in unregulated markets—almost as a matter of classifying these markets as not regulated—so that the relative regulated prices would be higher than could be explained by the δ and e_1 terms in the coefficient for P_2 . That is, uncertainty has the greatest effects in unregulated markets, and these effects are manifest in regulated prices higher than justified by regulatory and demand conditions.

Regulation while continuous could be uncertain, given that the profits allowed by case decision and administrative judgment have varied over time. The effect should appear in the estimated values of total profits M ; because regulation could be either more or less severe, M would be subject to stochastic variation. The relative regulated prices would be uncertain in the same way as costs or demands were uncertain. The results would be the same: the pipeline company can be expected to act as if M were lower than actually was the case under present regulation. Given this state of affairs, rather than construct more complex analyses with uncertainty introduced explicitly, it is preferable to evaluate the results from the

3. Data for testing regulated against unregulated prices

■ The information on prices charged by the interstate pipelines is voluminous, complicated, and not altogether useful for testing the effects of regulation. The companies file annual reports with the FPC on sales revenues and price schedules or "tariffs" for each contract over which the Commission has jurisdiction.¹⁷ The companies also file reports on sales revenues (but not tariffs) on unregulated transactions. Similarly information on physical volumes is extensive and differs between regulated and unregulated transactions. The Commission requires annual accounts of gas delivery under regulated contracts by tariff, buyer, and location of delivery. Information on the unregulated sales is limited to the amount of gas and the identity of the buyer.

These data are not altogether useful because they do not show demand charges, commodity charges, and volume by contract. The appropriate price information can be approximated for regulated sales by collating the annual volume by tariff with the marginal or commodity price shown on the tariff sheets; this has been done for the 1969 contracts of seven pipelines.¹⁸ The prices for unregulated sales are "average prices," calculated by dividing the total revenues from each industrial sale by physical volumes of deliveries on those sales,¹⁹ for each buyer at each location for these seven pipelines in 1969.

The price information is used to test coefficients a and b in $\{P_1 = aP_2 + b[mct(q_1) - mct(q_2)]\}$ for the effects of pipeline demand conditions, transport costs, and for the effects of regulation. The tests call for calculation of a and b in least squares regression equations, given data on P_1 , P_2 , $mct(q_1)$, and $mct(q_2)$. Since the equation is to be derived with respect to marginal prices, the "marginal or commodity price" on regulated sales is designated to serve as P_1 and the "average price" on unregulated sales is designated as P_2 .²⁰ Data are required on the marginal costs of transport, $mct(q_1)$ and $mct(q_2)$; these data are put together in the next section. After finding a and b , there are tests for estimating that portion of these two coefficients to be attributed to the value of δ , so as to infer the severity of regulation. The tests require independent estimates of the elasticities

certainty analysis in terms of whether some portion can be attributed to uncertainty.

¹⁷ Cf. the Form 2 Annual Reports of the Pipelines to the Federal Power Commission.

¹⁸ The seven pipelines were Arkansas-Louisiana, Cities Service, Panhandle Eastern, Southern Natural, Transcontinental, Tenneco, and United Gas Pipeline.

¹⁹ The treatment here of average revenue as the commodity price, with zero demand price, is justified by general practice with regard to demand prices on industrial sales. Interviews with executives of a number of the larger pipelines have led to the general conclusion of an absence of such demand prices.

²⁰ In a few cases, there may have been a demand charge and a commodity charge on q_2 . With $P_2^* = P_2 + w/q_2$ as the unit price, where w is the "demand" additive to the marginal or commodity charge, the equation is still the same—that is, $\{P_1 = aP_2^* + b[mct(q_1) - mct(q_2)] + d\}$ with a and b as given above—but with the addition of a constant term d . The equation can be derived by substituting $P_2^* = P_2 + w/q_2$ in the appropriate objective functions for maximizing profits, and then finding marginal pricing conditions as in the preceding section of this chapter. If $\partial w/\partial q_2 \neq 0$, or charges in fact do vary with the amount of gas delivered, then an additional term appears. This addition term is a constant denoted " d ."

of demand, e_1 and e_2 ;²¹ these estimates are made after the calculations of transport costs.

□ **The marginal costs of transport.** These costs can take on many different values, and finding the proper value depends on choosing the relevant time period, on accounting for the location of the pipeline, and on accounting for differences in factor prices of capital and fuel. For purposes of examining the results of recent price regulation, the time period of a single year in the interval 1952–1967 has been chosen and costs have been designated as “expenditures on operations, maintenance, administration, transmission, storage and depreciation” in each of these years for each of eight large interstate pipeline companies.²² These companies have been chosen because they have been built over similar terrain with similar technology in their main lines; also, they had much the same system length and capacity, with more than 1,000 miles of total main line pipeline, 75 percent of which had diameter equal to or greater than 15 inches.²³

The most general equation form chosen for costs on these systems was $C = \alpha + \beta m \cdot q$, with m equal to mileage and q equal to the annual throughput volume of gas. The length of the main line of each transporter was estimated by inspection of maps and engineering drawings to provide data for the variable m , and throughput of gas from the physical volume data was recorded as q .²⁴ When fitted to the sample, the equation proved to be $\{C = 44.7(10^3) + 0.57m \cdot q\}$, $R^2 = 0.958$ with α and β in cents per thousand cubic feet of gas transported for 100 miles. Attempts were made to account for differences in input factor prices among the transporters, and for the “mix” of volume and mileage as well, but they did not improve the fit of this equation (as measured by R^2). The introduction of variables to account for the year of the observation and/or to account for the particular pipeline did not change the values of α or β nor increase the value of R^2 . Thus the measure to be used here of marginal costs (mct) is 0.57 cents per hundred miles, so that values of transport cost differences $mct(q_1) - mct(q_2)$ are taken to be $0.57(m_1 - m_2)$ where m_1 and m_2 are 100 mainline miles of distance from a common gas field origin point to the points of resale of regulated gas q_1 and unregulated gas q_2 .²⁵

²¹ Given, for example, $a = \delta \frac{(1 + 1/e_1)}{(1 + 1/e_2)}$ then to find the regulatory effect

$\delta = \frac{a(1 + 1/e_2)}{(1 + 1/e_1)}$, estimates of a , e_1 , and e_2 are required.

²² See [14].

²³ The pipelines were: The Atlantic Seaboard Corporation, Colorado Interstate Natural Gas Pipeline, Michigan-Wisconsin Pipeline Company, The Natural Gas Pipeline Company of American, Tennessee Gas Transmission Company (a division of Tenneco), Texas Eastern Transmission Company, Transcontinental Gas Pipeline Corporation, and Trunkline Gas Corporation. The exception to the rule on location similarity is the Colorado Interstate Company.

²⁴ Volume is recorded as annual sales in [14].

²⁵ This approximation to cost implies that marginal costs do not change with respect to p , so that $\lambda \leq 1$ and $\delta \geq 1$ are the admissible values for the effects of regulation. The one possible case of lower relative prices under regulation $\delta \leq 0$ —when marginal transport costs are rising—has been excluded by this empirical approach.

□ **Elasticities of market demand.** The price sensitivity of gas demands has been estimated in a number of recent studies, but either for final sales to consumers or for sales at the well-head (so as to analyze the effects of regulation of field producer prices).²⁶ There have been no studies of the specific elasticities of demand of those buying gas at wholesale from the interstate pipelines. In order to make first estimates of e_1 and e_2 , here a sample of contract prices and quantities has been constructed and a regression analysis of the sample completed with equation forms suggested by the more detailed studies.²⁷

In most cases, markets for gas in which the pipelines are the sources of supply encompass whole metropolitan regions. The purchasing retail distributors can go to alternative sources of gas that deliver within their metropolitan regions, whether pipeline or LNG company; the industrial consumer can go to the same sources in the same metropolitan regions, as well as suppliers able to provide large volumes of coal and fuel oil. Data have been collected on transactions in a sample of 17 such market regions—specifically on the 1969 home fuel prices, on the gas tariffs and the industrial “average prices,” and on the contract volumes of gas delivered to both regulated and unregulated buyers.²⁸

For retail utility demand, functions have been estimated of the form

$$\log(q_1) = \alpha + \beta \log(P_1) + \gamma \log(P_F) + \delta \log(Y) \\ + \epsilon \log(N) + \phi \log(TDD) + u,$$

where q_1 is annual gas delivery by pipeline to the buying retail utility, P_1 the commodity price to the utility, P_F the price index of alternative home heating fuels, and Y , N , TDD the “market size” variables per capita income, population, and temperature degree days, respectively.²⁹ The distributor demand function proved to be

$$\log(q_1) = 8.106 - 1.907 \log(P_1) + 0.390 \log(P_F) \\ (1.342) \quad (0.816) \\ + 0.545 \log(Y) + 0.564 \log(N) - 0.445 \log(TDD); \quad (12) \\ (2.792) \quad (0.337) \quad (0.711)$$

$$R^2 = 0.374,$$

a relationship in general agreement with empirical regularities found elsewhere in energy demand studies. Gas price increases had the effect of decreasing quantity demanded, and income, population and other fuel price increases had the effect of increasing demand. But the regularities are not extremely marked—the coefficients are not

²⁶ Cf. Balestra and Nerlove [3] and MacAvoy [9].

²⁷ See Balestra [2] and Houthakker and Taylor [5].

²⁸ The observations on regulated sales conform to “market transactions” in the sense that they are each coincident with that from a metropolitan region. But information on the 49 purchases by industrial buyers was more limited, since a number of them took place in the same metropolitan market regions. Also, prices for alternative industrial fuels were not obtained in sufficient detail to be useful. Therefore, observations on industrial sales are more limited than they should be to indicate market behavior—they do not cover all the volumes within a market region containing the specific sale, nor the prices of alternatives.

²⁹ In the absence of more than one year’s data, lagged terms in q and P_1 to account for adjustment processes had to be deleted.

statistically significant³⁰ and the equation explains only 37 percent of the variation in the quantity variable. Although the price elasticity seems quite high—the coefficient equal to the e_1 estimate is -1.907 , clearly in the elastic range—it is not statistically significantly different from zero. The estimate also seems to be subject to wide variations; after collecting 1968 data on q_1 and N for these 17 locations, the regression was refitted for $q_1 = f[q(t-1), P, N, \dots]$ with the result that the coefficient for $\log(P_1)$ equaled -0.950 and for $\log[q(t-1)]$ equaled $+0.970$; both were statistically significant and $R^2 = 0.998$. The coefficient on $\log(P_1)$ is an estimate of the short-run price elasticity, which is closer to zero than the long-run elasticity. Thus, the two equations suggest, but do not establish, that demand is more likely to be elastic in the long run and inelastic in the short run.

The 1969 data on industrial demand, including price and quantity information on 49 direct pipeline sales, have been used to fit a least squares regression equation. Using corporate employees (EES) as a measure of “size of market,” the industrial demand regression was found to be:

$$\begin{aligned} \log(q_2) = & 11.372 - 1.785 \log(P_2) - 0.249 \log(dP32) \\ & \quad (0.980) \quad (0.114) \\ & - 0.114 \log(dP33) - 0.205 \log(dP29) + 0.681 \log(EES); \quad (13) \\ & \quad (0.125) \quad (0.168) \quad (0.192) \\ & R^2 = 0.306, \end{aligned}$$

where $dP32$, $dP33$, and $dP29$ are dummy variables for price additives on sales to three industries. The coefficients for price P_2 and for number of employees EES are statistically significant, and the first of the price additives (for classification 32, the stone-clay-glass industry) is also different from zero. The remaining coefficients do not differ from zero, and R^2 indicates that only 30 percent of the variance in q_2 is explained by the regression. But demands would appear to be quite elastic. The elasticity for the base industry (chemicals) is -1.785 , for industry 32 (stone-clay-glass) is $-1.785 - 0.249 = -2.034$, for industry 33 (primary metals) is $-1.785 - 0.144 = -1.929$, and for industry 29 (petroleum refining) is $-1.785 - 0.205 = -1.990$.

Further regression analysis suggests that industrial demand is even more elastic than these estimates indicate, particularly where there are close gas substitutes. Data were available for a more limited sample of 39 sales contracts with buying companies in the chemicals, glass, and primary metals industries, including price P_2 , quantity q_2 , and also including industrial coal prices P_c at the locations of gas delivery. The regression from this sample was:

$$\begin{aligned} \log(q_2) = & 10.954 - 3.471 \log(P_2) - 0.962 \log(dP32) \\ & \quad (0.918) \quad (0.542) \\ & - 1.070 \log(dP33) + 1.578 \log(P_c); \quad R^2 = 0.378. \quad (14) \\ & \quad (0.673) \quad (1.346) \end{aligned}$$

Here the estimated elasticities varied from -3.47 (on sales to firms in the chemicals industry) to -4.54 (on sales to companies in the

³⁰ The standard errors of the regression coefficients are shown in parentheses below the coefficients.

primary metals industry). Gas demand would seem to have been less elastic where the use of gas was as a process raw material with only fair substitutes (as in chemicals) and more elastic where the use was as a boiler fuel with very close substitutes (as in metals).³¹

Equations have also been fitted for two of the industries separately where there were sufficient data. These show substantial price elasticities and also indicate substantial differences in particular industries' reactions to gas price changes.³² The equation for demand for gas in the chemicals industry is

$$\log(q_2) = -2.021 - 3.486 \log(P_2) + 0.839 \log(EES) + 4.049 \log(P_c); \quad R^2 = 0.567, \quad (15)$$

(2.118) (0.315) (1.938)

for 12 observations, and the equation for demand in the stone-clay-glass industry is

$$\log(q_2) = 11.641 - 2.778 \log(P_2) + 0.819 \log(EES) + 1.163 \log(P_c); \quad R^2 = 0.232, \quad (16)$$

(1.371) (0.239) (0.998)

for 19 observations. They indicate values of price elasticity equal to -3.49 and -2.78 , respectively, and as such create the impression that industrial demand is more elastic than distributor demand. Since the industrial sales observations do not encompass whole markets but rather only individual transactions—so that they understate quantity responsiveness to price—the market elasticities have probably been greater. In the calculations for the δ effects from regulation, we shall assume that the elasticity of demand for distributor services lies between -1.0 and -2.0 and for industrial sales varies between -2.0 and -3.5 , which covers the point estimates from each of the regression equations.

□ **Elasticities of demand facing the individual pipeline.** So far we have estimated the “market elasticity” of demand—the demand schedule that a monopoly seller of gas would face in each market. Such a schedule measures the effect of competition from other fuels on the gas sellers' price, but it does not measure the effect of competition from other pipelines. In fact, we are interested in the demand schedule that faces the individual pipeline. Such a schedule will reflect *any* competition—both among fuels and among pipelines—that tends to cause a pipeline to lose sales as it raises prices or to gain them as it lowers prices. Rivalry among gas pipelines, in other words, might lead the individual pipeline to set prices that are lower than the

³¹ A more complete but not necessarily illuminating form is as follows:

$$\log(q_2/EES) = 0.757 - 2.168 \log(P_2/P_F) - 1.109 \log(dP32) - 1.049 \log(dP33) + 2.958 \log(P_c/P_F); \quad R^2 = 0.364, \quad (14a)$$

(1.175) (0.542) (0.364) (0.821)

where P_c is price of coal and P_F is the retail fuel price index.

³² These industry differences may explain the low values of R^2 in the combined-industries equations.

profit-maximizing level that our measure of “market elasticity” implies.

Rivalry among pipelines serving the same industrial buyers, or even the same retail utility, could by itself have lowered prices more than the actions of the Commission. The additional element of separate pipeline price initiatives can be taken account of by modifying the e_1 and e_2 calculations so as to include “Cournot” elements. The “Cournot” argument that each firm sets its own sales and prices to increase its market share replaces e_i by ne_i for n equivalent-sized firms and for e_i as market-wide price elasticity.³³ It is important to assess the extent of such individual price-reducing initiatives, in terms of finding ne_2 rather than e_2 in the relative price equation $P_1 = aP_2 + b[mct(q_1) - mct(q_2)]$.

Consider the relative price equation

$$P_2(x) = aP_2(y) + b[mct(q_{2x}) - mct(q_{2y})],$$

where $P_2(x)$ and $P_2(y)$ are both industrial prices, one for sales to industry x and the other to industry y . This equation can be fitted by least squares regression to the data from the seven pipelines’ industrial sales in 1969. According to the analysis in the last section, the calculated $a = \hat{a} = (1 + 1/e_{2y})/(1 + 1/e_{2x})$ and the calculated $\hat{b} = 1/(1 + 1/e_{2x})$ since both sets of industrial sales prices were unregulated.³⁴ If there were systematic Cournot rivalry in the 1960’s, then $\hat{a}$ and $\hat{b}$ would have been too low to be “explained” by the values of e_{2x} and e_{2y} found above, and these values of e_{2x} and e_{2y} should be replaced by ne_{2x} and ne_{2y} with “ n ” as “observed” in this new equation.³⁵

This equation has been fitted by least squares to 40 observations of 1969 annual sales of four pipelines to companies in the petroleum industry and the primary metals industry. The prices are $P_2(29)$ for petroleum and $P_2(33)$ for primary metals, the marginal costs are $mct[q_2(29)]$ and $mct[q_2(33)]$ for transporting along any one of the four pipelines. The relationship in the regression is:³⁶

$$P_2(29) - P_2(33) = \Delta P_2 = (a - 1)P_2(33) + b\{mct[q_2(29)] - mct[q_2(33)]\} + d. \quad (17)$$

The first comparative equation for price is:

$$\begin{aligned} \Delta P_2 = & -0.936P_2(33) \\ & (0.048) \\ & + 1.301\{mct[q_2(29)] - mct[q_2(33)]\} + 26.64; \quad R^2 = 0.947, \\ & (0.182) \end{aligned} \quad (18)$$

with all prices in cents per *Mcf* of gas, and the implied elasticity of demand of buyers in petroleum $ne = -4.32$.³⁷ The second comparative price equation is for 87 observations on $P_2(33)$, primary

³³ Cf. Vickrey [15], pp. 337–338.

³⁴ That is, without regulation, $\delta = 1.0$ in this equation.

³⁵ That is, with $\hat{b} = 1/(1 + 1/ne_{2x})$, then $n = 1/(\hat{e}_{2x}[(1/\hat{b}) - 1])$.

³⁶ The constant term accounts for pipeline-to-pipeline differences in transport costs. These differences are ignored in $mct(q_2)$ by taking $0.57(m_1 - m_2)$ for all lines, with m_1 and m_2 hundreds of miles at 0.57 cents per hundred miles.

³⁷ The value for ne is calculated from

$$\hat{b} = 1.301 = \frac{1}{(1 + 1/ne)} \quad \text{or} \quad ne = \left(\frac{1}{1.301} - 1 \right)^{-1}.$$

metals, against $P_2(49)$ electricity:

$$\Delta P_2 = -0.572P_2(49) + 2.581\{mct[q_2(33)] - mct[q_2(49)]\} + 19.86; \\ (0.603) \quad (0.572) \quad R^2 = 0.245, \quad (19)$$

and the implied elasticity of demand in primary metals $ne = -1.63$. The third comparative price equation is based on 242 observations of contracts with stone-clay-glass manufacturers (32) and the electricity generating companies (49):

$$\Delta P_2 = -0.407P_2(49) + 1.061\{mct[q_2(32)] - mct[q_2(49)]\} + 13.59; \\ (0.064) \quad (0.298) \quad R^2 = 0.161, \quad (20)$$

with the elasticity of demand of stone-clay-glass manufacturers $ne = -17.66$. The fourth comparative price equation, from 44 observations of $P_2(28)$ chemicals against $P_2(33)$ primary metals, is:

$$\Delta P_2 = -0.927P_2(33) \\ (0.120) \\ + 1.504\{mct[q_2(28)] - mct[q_2(33)]\} + 34.99; \quad R^2 = 0.735, \quad (21) \\ (0.426)$$

and the implied elasticity of demand ne for chemicals is -2.98 . The four equations show estimates of ne that are larger in half the cases than those for market demand e given above; two (petroleum and stone) showed much larger values of ne and the other two showed approximately the same values for ne and e . In general, given varying quality of industry sales data, it has to be concluded that n was equal to or greater than one. That is, the elasticity of demand for the purposes of individual pipeline price setting was greater than the industry-wide demand elasticity. In at least some cases the presence of more than one pipeline apparently had a price-reducing effect.

The price-setting practices in markets for pipeline sales to retail gas utilities seem to have been quite different from those in industrial markets. The pipelines set two-part prices on contracts with retail utility companies. The marginal or commodity prices have not been set at a level in keeping with the residential market elasticities of demand, but as if they were being used to add to marginal consumption while the monthly or demand prices accumulated the bulk of the net revenues from providing service. This can be seen from a sample of 26 paired residential sales (paired within the same pipelines for 1969) fitted to the regression³⁸ $\Delta P = b[mct(q_{1x}) - mct(q_{1y})]$ for sales at locations x and y . The equation is $\Delta P = +0.869[mct(q_{1x}) - mct(q_{1y})]$ and $R^2 = 0.312$, with the fitted coefficient $\hat{b}$ less than 1.0. The commodity prices at farther locations did not compensate for marginal transport costs to those locations. A second regression was fitted as follows: $\Delta AR = 2.451[mct(q_{1x}) - mct(q_{1y})]$ with $R^2 = 0.317$, for AR equal to the "average price" from averaging demand plus commodity charges over the volumes of gas delivered; here the elasticity ne implied by the fitted value of $\hat{b}$ is -1.70 . The ΔP regression indicates that commodity prices have not been set with

³⁸ Here $\{P_1 - P_2 = \Delta P = (a - 1)P_2 + b[mct(q_1) - mct(q_2)] + d\}$ with $a = 1$ and $d = 0$, because the samples q_1 and q_2 both have the same price elasticity and are from the same pipelines.

respect to market demand elasticity,³⁹ but the ΔAR regression indicates that the average of demand and commodity prices has been set close to the equivalent "best profit" single charge if market demand elasticity were close to -1.70 . Since the value of -1.70 is within the range of elasticities observed for retail gas utility markets in 1969, it has to be concluded that ne equals the market elasticity e . Demand and commodity prices on regulated sales reflect little or no rivalry among the individual pipelines.

On the whole, prices for gas to industries seem to have been set with regard to individual seller demand elasticities ne_2 higher than market elasticities e_2 , in keeping with some pipeline rivalry. The number of pipelines at any industrial consumption location has made a difference for the level of prices there—for price setting, the number of independent suppliers n is between 1 and 2, and price differences have been reduced by somewhat less than one-half of the single seller level because of rivalry among these suppliers.

In contrast, prices have been set for retail utilities in two parts, with the marginal price designed to induce sales at or somewhat below marginal costs, without regard for individual seller demand elasticity.⁴⁰ The initial price, making up the second part of the tariff, has been "in line" with market demand elasticity. The regulated tariffs then effectively followed the design for monopoly and discriminatory pricing.⁴¹

This is important for the analysis of regulation, since it leads to two conditions. First, the industrial elasticities contained in a and b in the relative price equations are within the range -3.0 to -5.0 . Second, the retail gas utility elasticities contained in a and b are effectively eliminated from the analysis, since only commodity prices of two-part tariffs are used, and the commodity charges imply no conformity to retail market elasticities.⁴²

³⁹ In fact, the commodity prices are more similar to the model for perfectly discriminatory prices (when all pipelines operate together to control the market for maximum joint profits) than to any other model used here. But the price differences are too small to have been perfectly discriminatory—they were less than the calculated differences in costs of transport. This behavior is rational when marginal buyers are large relative to the intramarginal (closer) buyers. Cf. Oi [12] for the tradeoff between changes in demand and commodity prices along these lines.

This behavior also indicates that the primary rationing mechanism for these companies has been the commodity price. If regulation were effective, then, in reducing profits, it would operate to increase the commodity price and release capacity for industrial use.

⁴⁰ It is profitable to go below marginal costs on the commodity price if the demand price can be set for that class and for other buyers so as to make up more than the marginal loss; cf. Oi [12].

⁴¹ *Ibid.*, pp. 84–85. For the retail utility sales, the term in b for $(1 + 1/e_1)$ should be replaced by $1 + (1 - Ns_1)/e_1$, where s_1 is the market share of the marginal consumer at the end of the line and N is the number of consumers. With the marginal consumers consisting of large retail buyers in the northern population centers, $1 - Ns_1 < 0$ and it is profitable to increase commodity prices at the farther locations by less than the marginal costs of additional transport.

⁴² The elimination of the elasticity e_1 is implied by perfect discrimination—by demand charges tailored to each market and differential commodity prices equal to differential costs of transportation, as in equation (2) where $P_1 = P_2 + (\partial C/\partial q_1 - \partial C/\partial q_2)$. The more limited two-part tariff, where demand and commodity prices to all within the retail utility class at one site have to be uniform, has similar implications if the term replacing $1/e_1$ —the term

4. Testing regulated against unregulated prices

■ Finding the extent to which profit or rate-of-return regulation affected prices in the 1960s now requires calculation of the coefficients a and b in $\{P_1 = aP_2 + b[mct(q_1) - mct(q_2)]\}$. Here the sample of 1969 contracts for seven pipelines with industrial and retail utility buyers is used to fit a regression of regulated marginal price P_1 on unregulated unit price P_2 , the difference in costs of transport $[mct(q_1) - mct(q_2)]$, and a constant. In this least squares equation

$$P_1 = aP_2 + b[mct(q_1) - mct(q_2)] + d_1L_1 + d_2L_2 + \dots + d_6L_6. \quad (22)$$

The constant term is a series of dummy variables, $L_1, \dots, L_6$, taking on values of "1" for six of the seven pipelines (to account for the effects of differing field gas purchase prices on the absolute levels of resale regulated prices).⁴³ The observations make a sample of 986 paired residential/industrial transactions and the least squares regression from this sample is:⁴⁴

$$P_1 = 0.906P_2 + 0.504[mct(q_1) - mct(q_2)] - 2.531L_1 - 4.617L_2 - 13.625L_3 - 8.300L_4 - 8.500L_5 + 8.801L_6. \quad (22a)$$

The relationship indicates that prices were lower on regulated sales, and less-than-compensatory price markups were taken on increased transport costs to regulated sales at farther locations. Once account is taken of differences in demand elasticities, however, the regulated prices were probably not greatly different from unregulated prices with the same demand conditions.

With the estimated value of $\hat{a} = 0.906$, and with

$$\hat{a} = \delta[(1 + 1/e_2)/(1 + 1/e_1)],$$

we can insert probable values of the coefficients of elasticity e_1 and e_2 and thereby derive δ . The elasticity of industrial demand e_2 takes on values from -2.5 to -5.5 , and for obtaining all "reasonable" values of δ , it is assumed that values of e_2 of -3.0 , -4.0 , and -5.0 occur with probability 0.3 and values of -2.5 and -5.5 occur with

$(1 - N_{s1})/e_1$ —is zero. This occurs when all retail buyers are the same size or the marginal buyer is at least of the same scale as the intramarginal (closer) buyers. In this set of circumstances, the results for relative prices is the same as if e_1 were infinite. The circumstances would appear to be descriptive of those in pipeline sales, where the buyers at the farthest locations are not only larger, but have relatively easy access to other fuels (as do cities in the North and East).

⁴³ The logic of this can be seen from separating pipeline costs into (a) field purchase price, and (b) transmission costs, and then deriving first order conditions for maximum profit resale prices. The field purchase price enters as the constant in the equation $\{P_1 = aP_2 + b[mct(q_1) - mct(q_2)] + d\}$. This constant will differ for each pipeline. This dummy will also pick up the effect of interregional differences in average costs, which may be important in determining relative prices (see note 12).

⁴⁴ This regression was fitted with P_2 as the dependent variable as follows:

$$P_2 = 1.104P_1 - 0.556[mct(q_1) - mct(q_2)] + 2.794L_1 + 5.098L_2 + 15.042L_3 + 9.163L_4 + 9.384L_5 - 9.717L_6; \quad R^2 = 0.551. \quad (22b)$$

(0.008) (0.111) (0.248) (0.714) (0.535)
(0.705) (0.819) (1.003)

The reason for the reversal of variables was that more serious errors in data were expected on unregulated sales, and these could better be relegated to the dependent variable.

probability 0.05 each.⁴⁵ The elasticity of retail utility demand e_1 is eliminated from this calculation because of the structure and nature of two-part prices on sales to utility buyers. There is variation in $\hat{a}$ as well; the standard error of $1/\hat{a}$ is 0.008 (as shown in footnote 44). Taking five values of e_1 and twenty values of $\hat{a}$ (as given by a normal distribution with standard deviation of 0.008), the values of δ range from 1.475 to 1.109 and the average value of δ equals 1.206.⁴⁶ These values of $\hat{\delta}$ indicate that the Federal Power Commission profit controls had some effect. Relative prices on regulated sales were increased by the δ ratio 1.2, whereas relative prices without regulation would have had the δ ratio 1.0.

The effect need not have been substantial, however. The effects of uncertainty only could account for a value of $\delta \geq 1.2$. Stringent regulation would have produced values of δ from five to ten times larger.⁴⁷ Within the range of "reasonable market demand conditions," as shown by the chosen values of e_1 and e_2 , and the calculated range of $\hat{a}$, the effects from regulation were very moderately price controlling.⁴⁸

The other parts of the regression equation are also revealing of the price-setting process. The coefficient b for $[mct(q_1) - mct(q_2)]$ has low values in the estimated equation—there seems to have been much less price increase with greater distance than called for by transport cost differences. This can be at least partly explained by examination of the data: most of the actual sales at the "farther distance" have been to retail public utilities, and the relatively lower commodity

⁴⁵ The range from -2.5 to -5.5 encompasses values of estimate $-l_2$ and ne_2 both, for n between 1 and 2.

⁴⁶ The end values of the range of δ are given with a probability of 0.05.

⁴⁷ With $\delta = 1/(1 - \lambda)$ and $\lambda \rightarrow 1.0$ as regulation increases in stringency, $\delta \rightarrow \infty$.

⁴⁸ The price differences between retail utility buyers and particular buying industries were similar to those shown by the pooled industry-retail sample. But the "retail-particular industry" differences show some effects from regulation specific to pricing patterns on the unregulated industry sales. For unregulated prices to buyers in the chemicals and stone-clay-glass industries, versus regulated prices, the regression equations had $1/\hat{a}$ coefficients equal to 1.323 and 1.365 respectively, which imply values of δ lower than those derived from the all-industries regression; here $\delta = +0.96$ for $e_2 = -3.0$, with a range from 0.80 to 1.25 for retail utility-chemicals and $\delta = 0.77$ with a range from 0.70 to 0.85 for retail utility versus stone-clay-glass. There would seem to have been no effects whatever from regulation of prices to retailers, when these are compared to prices on industrial sales in the chemical and stone-clay-glass industries, since values of δ less than 1.0 indicate no change in profitability from regulation. The equations for direct pipeline buyers in the petroleum and metals industries versus retail utility buyers indicate a greater effect. The calculated values of $1/\hat{a}$ were 0.756 and 1.545 and, given the e_2 specific to these industries, these two regressions show stronger δ effects from regulation. The coefficients for P_1 imply values of δ from 1.33 to 2.30 for retail utility versus petroleum and from 1.19 to 6.25 for sales to retailers versus primary metals. These values can be interpreted to imply that more of the results from "substantial regulation" were realized here.

There is no reason, however, for regulation to appear to be more stringent from one sample regression to another, given that the only differences between samples are in the unregulated industries with which comparisons of regulated prices are being made. Rather, the rationale might be that, in particular industry regressions, we are observing price variations having to do with particular competitive conditions in these samples of unregulated sales—conditions not directly incorporated in measures of elasticity of demand. This rationale takes away credit given to "regulation" that is due to "competition" in the unregulated markets.

prices on these sales have been confounded with the "distance" effect.⁴⁹ The two-part prices on the regulated utility sales have been set so that the commodity prices P_1 discriminate in favor of buyers at farther locations (as has been seen above). With a predominance of these prices at greater distance " m ," the paired transactions show the effects in $\hat{b}$ of discrimination within regulated sales only.⁵⁰

Thus the process of regulatory control might be viewed here as having taken place in two steps. First, price setting under regulation has involved the intensive use of two-part prices, with initial charges on monthly service and marginal prices per *mcf* of delivered gas. The marginal regulated prices were substantially lower than "average prices" on unregulated sales even though they may have actually been more profitable (as one part of two-part prices). Second, regulation by reducing profitability has raised that marginal price above the level that it would have had without regulation. This second step, the relative price effect from regulation, was of modest size.

Regulation, by providing the framework for two-part pricing, might be credited with reduced marginal prices at farther locations. There might have been as much rivalry on these transactions, in the absence of the Commission process, as on unregulated transactions—with the result that there would have been no two-part tariffs. But did the practice of such discrimination under regulation imply a lower price level for those consumers? This depends on the extent of the separate effects of profit regulation. Estimation of this effect alone is not possible, because the two-part prices were the only prices available—there was no set of single-price regulated sales with which to compare single-price unregulated sales. Only the crudest indication can be found of the effects of regulation alone—from comparing *average revenue* AR_1 , the average of demand and commodity prices on regulated sales, with the unregulated price P_2 . The "price" AR_1 can be considered the "equivalent" one-part tariff, given that the FPC had set rates of return on capital so that the profits were the same as were collected under the two-part tariffs. The equation $\{AR_1 = aP_2 + b[mct(q_1) - mct(q_2)] + d\}$ then shows the relative price effects from regulation in the a and b coefficients.

This equation has been fitted to the 986-observation sample of pooled industrial-retail utility sales in 1969. The resulting least squares regression line is as follows:

$$AR_1 = 1.086P_2 + 0.739[mct(q_1) - mct(q_2)] - 2.931L_1 - 14.043L_2 - 3.362L_3 - 5.937L_4 - 2.094L_5 + 0.739L_6, \quad (23a)$$

indicating higher relative prices for regulated sales than shown in the

⁴⁹ Another explanation could be that the value of γ , the portion of costs allocated to regulated sales, has been very high; but since $b = [(1 - \gamma\lambda)/(1 - \lambda)] \times [1/(1 + 1/e_1)]$, with $e_1 = \infty$ and λ estimated from a , the estimated value of γ would have to have been greater than 100 percent. This would be to suggest that all costs have been allocated to regulated sales—not a likely interpretation of present practice.

⁵⁰ There are no further tests possible with the L_i coefficients. The different pipelines have had different price levels for retail utility sales, perhaps because of different degrees of regulatory control applied by the FPC, but more likely because of historical variation in the prices paid for field supplies of gas. Earlier pipelines paid lower prices for gas supplies than pipelines constructed later, and, with these earlier prices still in effect, the final sales prices of the earlier pipelines have been lower.

marginal price comparisons.⁵¹ But this relative price ratio $\hat{a}$ has to be separated into its component parts based on e_1 and e_2 to find the regulatory ratio δ . The values of e_2 range from -2.5 to -5.5 , as assumed above. The values of the elasticity e_1 without a two-part tariff lie in the range from -1.0 to -2.0 (with -1.5 assumed to be most probable). Taking combinations of e_1 and e_2 within these ranges and $\hat{a}$ within one standard deviation of the calculated 1.086, the estimated values of δ are less than one, ranging from 0.15 to 0.80. Generally, these values of δ are not consistent with $\delta \geq 1$ expected from effective regulation with the positive cost and demand conditions found here. Rather, these estimated values of δ are in the range indicating no effects whatsoever from regulatory constraint on pipeline company rates of return.⁵²

■ This conclusion must be brought together with those from the comparisons of marginal prices, and those from the earlier price equations on industrial sales free of the effects of regulation. The assessment of findings from the equation for average price AR has to be quite tentative, since it is not possible to define precisely what is being measured by "price" in terms of the average revenue from a two-part tariff. But if equal weight were to be given to the AR and the commodity price regressions, the effects of regulation would be shown by calculated values of δ of 1.2 (using P_1 and P_2) and of 0.8 (using AR_1 and P_2). The first value implies some effect from regulation—the constraint "binds" so that a \$1.00 increase in *allowed* profit returns would increase *actual* profits by \$.16—but the lower value of δ implies that a \$1.00 increase in *allowed* returns would have no effect on *actual* returns, because actual returns were already at a maximum, given market conditions.⁵³ If less than equal weight were put on the AR regression, the conclusion would be that there has been a slight tendency towards higher relative prices under regulation, but not a strong tendency.

Prices charged by pipelines on regulated and unregulated sales increased with higher costs of transport, but not by as much as the increased costs. Prices have been higher on sales with lower elasticities of demand. Prices at the margin on regulated sales have been reduced by the use of two-part tariffs; but after accounting for the specific effects of the two-part tariff, the prices showed slight additional effects from profit controls. That is, after taking account of variation with costs, with demand elasticities and with tariff structure, regulation appears to have had little effect on prices.

⁵¹ The fitted equation was:

$$P_2 = 0.921AR_1 - 0.681[mct(q_1) - mct(q_2)] + 2.699L_1 + 12.934L_2 + 3.096L_3 \\ (0.008) \quad (0.132) \quad (0.297) \quad (0.645) \quad (0.865) \\ + 5.468L_4 + 1.929L_5 - 0.681L_6; \quad R^2 = 0.366. \quad (23b) \\ (0.983) \quad (1.162) \quad (0.132)$$

⁵² There are, of course, other alternatives than designating the calculated AR_1 as the regulated unit price. There is little chance that, given a regulatory profit constraint, the pipeline sellers of gas would charge the price exactly equal to the average revenue from the two-part tariffs.

⁵³ With $\delta = 1/(1 - \lambda)$ and the λ constraint equal to the change in actual with respect to allowed profits $\partial L/\partial M$, values of $\delta \sim 1.2$ are implied by $\lambda \sim 0.16$ or changes of 0.16 per dollar of increased *allowed* profits M .

5. Summary

References

1. AVERCH, H. AND JOHNSON, L. L. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol. 52, No. 5 (December 1962), pp. 1053-1069.
2. BALESTRA, P. *The Demand for Natural Gas in the United States*. New York: North-Holland Publishing Company, 1967.
3. ——— AND NERLOVE, M. "The Demand for Natural Gas." *Econometrica*, Vol. 34, No. 3 (July 1966), pp. 585-612.
4. BAUMOL, W. J. AND KLEVORICK, A. K. "Input Choices and Rate of Return Regulation: An Overview of the Discussion." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 162-190.
5. HOUTHAKKER, H. S. AND TAYLOR, L. D. *Consumer Demand in the United States, 1929-1970*. Cambridge: Harvard Univ. Press, 1966.
6. KAHN, A. E. *The Economics of Regulation*. New York: John Wiley & Sons, Inc., 1970.
7. LELAND, H. E. "The Theory of the Firm Facing Uncertain Demand." Report No. 24, Stanford Institute for Mathematics, 1970.
8. MACAVOY, P. W. "The Formal Work-Product of the Federal Power Commissioners." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 385-387.
9. ———. "The Regulation-Induced Shortage of Natural Gas." *Journal of Law and Economics*, Vol. 14, No. 1 (April 1971), pp. 167-199.
10. MCCALL, J. J. "Probabilistic Microeconomics." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 2 (Autumn 1971), pp. 403-433.
11. MILLS, E. S. "Uncertainty and Price Theory." *Quarterly Journal of Economics*, Vol. 73, No. 1 (February 1959), pp. 116-130.
12. OI, W. Y. "A Disneyland Dilemma: Two Part Tariffs." *Quarterly Journal of Economics*, Vol. 85, No. 1 (February 1971), pp. 77-96.
13. PRESSMAN, I. "A Mathematical Formulation of the Peak-Load Pricing Problem." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 304-327.
14. U. S. FEDERAL POWER COMMISSION. *Statistics of Natural Gas Pipeline Companies. 1952-1967*.
15. VICKREY, W. S. *Microstatics*. New York: Harcourt, Brace & World, Inc., 1964.
16. WELLISZ, S. "Regulation of Natural Gas Pipeline Companies: An Economic Analysis." *Journal of Political Economy*, Vol. 71, No. 1 (February 1963), p. 33.

Market structure, quality and durability

David Levhari

Associate Professor of Economics
The Hebrew University of Jerusalem

and

Yoram Peles

Lecturer
School of Business Administration
The Hebrew University of Jerusalem

This paper analyzes the effects of market structure—monopoly versus competition—on the quality and durability of goods. Also, it tries to find the impact of government regulation on these variables. The types of quality improvements discussed are: quality as pure substitute for quantity; quality which increases the demand for the good; and quality improvement which increases the durability of the good. In general, it is impossible to deduce that quality is independent of market structure. It depends on the cost structure. The paper shows that when quality is a substitute for quantity, both quality and quantity of the monopoly might fall short of those in the competitive market. Regulating only quality, or only quantity, may increase the monopoly misallocation of resources. In other types of quality improvements discussed, it may turn out that quality and durability may be better or worse in the monopolized industry than in the competitive one. Regulating only quality may improve the resource allocation but not eliminate the bias. Quantity regulation by itself may be sufficient for producing the optimal flow of services.

■ The effects of market structure¹ on the quality and durability of the goods produced have been under discussion for a long period of

1. Introduction

David Levhari received his Ph.D. from M.I.T. He is Associate Professor of Economics at The Hebrew University of Jerusalem. Professor Levhari is a fellow of the Econometric Society and associate editor of *Econometrica*.

Yoram Peles received his Ph.D. from the University of Chicago. He is Lecturer at The Hebrew University of Jerusalem, School of Business Administration.

¹ The comparison between a monopolized and a competitive industry can be interpreted in one of two ways:

(1) The monopoly is a cartel of many firms whose marginal cost curve is the horizontal summation of the individual firms' marginal cost curves, which in turn have no external economies. The industry marginal cost curve can be horizontal or positively sloped, in both market structures.

(2) We are not considering monopoly versus competitive industry literally. The relevant comparison is between a monopoly maximizing its benefit versus the same firm acting to maximize the social benefit.

We prefer the second interpretation. Hence, whenever we speak of the behavior of the competitive industry, we actually mean the firm's behavior in the interests of the public.

time. Recently, there has been some revival of these discussions.² In the present paper we analyze these effects and try to find the impact of government regulation on the variables. The question is not only whether we can improve welfare by proper control of durability or quality where it is not possible to control the volume of output, but whether volume control is sufficient to restore the misallocation of resources caused by the monopoly. In particular, we discuss some special forms of quality as pure substitute for quantity, quality improvement which increases the demand for the good, and quality improvement which increases the durability of the good. We show that, in general, it is impossible to deduce whether the monopoly has an adverse effect on the quality or durability of the good: the effect depends on the cost structure. We show that when quality is a substitute for quantity, quality and quantity of the monopoly might fall short of those in the competitive market. Partial regulation, that is, regulation of quality only (or quantity only), may increase the monopoly misallocation of resources.

In other cases, it may turn out that quality and durability may be better or worse in the monopolized industry than in the competitive one. Partial regulation of quality or durability may improve the resource allocation but not eliminate the bias. Quantity regulation by itself may be sufficient for producing the optimal flow of services.

2. Quality of goods,³ constant demand for services

■ Consider a product whose useful life lasts only one period of time. In each period of time there is a given demand for the services supplied by the product. If, for example, there is an improvement such that a new unit of the product can give twice as much of the same type of service as an old unit, people will readily pay twice as much per unit of a new product as they did for the old. Suppose the product is razor blades, and a unit of service is one shave. If a physical unit of a blade can be used for ten shaves, it is conceivable that the price will be twice the price of a blade that can only be used for five shaves. The total demand for shaving is given. Given the technical possibility of producing blades of different quality, i.e., blades that can supply a whole range of number of shaves, which quality will the monopoly choose, and which will the competitive industry choose? Notice that the foregoing is a long-run analysis. When to implement the new, improved product is a short-run analysis, which is outside the framework of this article.

Let q denote the quality of the product, where the base is $q = 1$. Thus, a blade that can be used for five shaves has $q = 5$. And let subscript m denote monopoly and subscript c denote competition.

P is the price of one unit of the product characterized by $q = 1$ (namely, P is the price of a shave). Hence, $P = f(xq)$ where x is the number of units produced. Note the symmetry of the effects of x and q on the price. We shall have the same decline in P by multiplying x or q by a constant. The total revenue of the monopoly producer is: Pxq or $xqf(xq)$ and the profits he has to maximize are $xqf(xq) - C(x, q)$ where C denotes cost. The two marginal conditions we

² See [2], [3], [4], and [5].

³ On the general condition for joint optimization of quality and price see Dorfman and Steiner in [1].

have are:

$$\begin{aligned}\frac{\partial C(x, q)}{\partial x} &= qf(xq) + f'(xq)xq^2 \\ &= qf(xq)[1 + f'(xq)xq/f(xq)]\end{aligned}\quad (1)$$

or

$$\frac{\partial C(x, q)}{\partial x} = qf(xq) \left[1 + \frac{1}{\eta} \right]. \quad (2)$$

Note that $qf(xq)$ is the price of one unit x with quality q . Dividing (2) by q , we obtain

$$\frac{1}{q} \frac{\partial C(x, q)}{\partial x} = f(xq) \left[1 + \frac{1}{\eta} \right] = P \left[1 + \frac{1}{\eta} \right]. \quad (3)$$

The optimum condition for q is given by

$$\frac{\partial C(x, q)}{\partial q} = xf(xq) + x^2qf'(xq). \quad (4)$$

Dividing by x , so as to have the marginal cost with respect to q per unit of quantity x , we obtain

$$\begin{aligned}\frac{1}{x} \frac{\partial C(x, q)}{\partial q} &= f(xq) + xqf'(xq) = P + xq \frac{dP}{d(xq)} \\ &= P \left(1 + xq \frac{dP}{Pd(xq)} \right) = P \left(1 + \frac{1}{\eta} \right).\end{aligned}\quad (5)$$

From equations (3) and (5), we obtain the relationship between the monopoly optimum condition for x and q as follows:

$$\frac{1}{x} \frac{\partial C(x, q)}{\partial q} = \frac{1}{q} \frac{\partial C(x, q)}{\partial x}, \quad (6)$$

or, equivalently,

$$x \frac{\partial C(x, q)}{\partial x} = q \frac{\partial C(x, q)}{\partial q} \quad (7)$$

or

$$\frac{\partial C(x, q)}{\partial x} \bigg/ \frac{\partial C(x, q)}{\partial q} = \frac{q}{x}. \quad (8)$$

The competitive market solution (or socially optimum operating industry) will set the marginal cost with respect to x equal to the price of a unit of x given the firm's quality q , or

$$\frac{\partial C(x, q)}{\partial x} = Pq. \quad (9)$$

The marginal cost with respect to q will, in a competitive situation, be equal to the extra revenue derived from increasing q by one unit (over the whole of the firm's quantity x), that is

$$\frac{\partial C(x, q)}{\partial q} = Px. \quad (10)$$

From equations (9) and (10), we obtain for a competitive firm

$$\frac{\partial C(x, q)}{\partial x} \bigg/ \frac{\partial C(x, q)}{\partial q} = q/x. \quad (11)$$

As we see, the condition for optimal production both for monopoly and competitive industries is, $\frac{\partial C}{\partial x} \bigg/ \frac{\partial C}{\partial q} = \frac{q}{x}$, or alternatively the partial elasticities of the cost function,

$$\epsilon_C x = \epsilon_C q. \quad (12)$$

If we assume separability

$$C(x, q) = h(q)b(x), \quad (13)$$

then condition (12) yields

$$\frac{b'(x)}{b(x)} \cdot x = \frac{h'(q)}{h(q)} \cdot q. \quad (14)$$

Thus at optimal production the elasticities of b with respect to x and h with respect to q should be equal. In the case where $\frac{b'(x)}{b(x)} x = \alpha$, i.e. $b(x) = x^\alpha$, we obtain

$$\frac{h'(q)}{h(q)} \cdot q = \alpha, \quad (15)$$

and quality determination is independent of x and thus of market structure. A special case discussed in the literature is that of constant returns to scale where $C(x, q) = x \cdot h(q)$ and $\frac{h'(q)}{h(q)} \cdot q = 1$, and therefore, q is determined independently of market structure. Notice that in the symmetric case $h(q) = aq^b$, x is determined independently of market structure.

If $\frac{1}{x} \frac{\partial C}{\partial q}$ depends on x (namely the marginal cost of improving the quality of one unit of product depends on how many units of x are produced), we can get $q_m > q_c$, just as for other types of quality or durability changes discussed in subsequent sections. Moreover, by replacing x with q and q with x , we can obtain solutions in which $x_m > x_c$ and $q_m < q_c$.

□ **Implications for policy.** Regulating the monopoly can be done by setting standards for x or q , or both. If the regulatory body knows the social optimum x and q , setting standards for both will no doubt force the monopoly to produce the social optimum x and q . However, setting standards for only one variable, x or q , will not bring optimum production. Suppose a standard is set for q which forces the monopoly to improve the quality to that in the competitive market. The monopoly is still free to change x , and it will do so. Now, best production calls for reducing the output of x . Before regulation was in force, the monopoly used the best combination of input to x and input to q to produce the revenue (xqP) from the amount of services sold xq ; when q is regulated, more of q and less of x have to be used to produce the same amount of revenue as before, and, since the best

combination was used before, the present combination is inferior, i.e., the costs of producing the previous revenue (xqP) are higher. In short, the monopoly output of services xq is lower, and the cost per unit of service is higher with quality regulations than without these regulations. The same holds for setting quantity standards alone, leaving the monopoly free to change quality.

The above analysis can be shown in a more formal way, but first note that if either $\frac{\partial C}{\partial x}$ or $\frac{\partial C}{\partial q}$ is a decreasing function, i.e., $C_{xx} < 0$ or $C_{qq} < 0$ (or one is decreasing faster than the other), then we have a corner solution. When it is worthwhile to produce x (or q), it is worthwhile to continue with it at the expense of q (or x) until we get to the corner. If $\frac{\partial C}{\partial x}$ and $\frac{\partial C}{\partial q}$ are both constant functions, i.e., the cost function is linear in x and q , and also either

$$\frac{\partial C(x, q)}{\partial x} \frac{1}{q} < \frac{\partial C(x, q)}{\partial q} \frac{1}{x}$$

or

$$\frac{\partial C(x, q)}{\partial x} \frac{1}{q} > \frac{\partial C(x, q)}{\partial q} \frac{1}{x},$$

then, once again, we obtain a corner solution. The following discussion excludes these corner solutions.

Let us denote the preregulation solution by $\hat{q}$, $\hat{x}$ and the post-regulation solution by $\bar{q}$, $\bar{x}$. Suppose a regulatory commission sets a standard for q such that $\bar{q} > \hat{q}$. The question is whether the total amount xq of services produced will increase as a result. We claim that $\bar{q}\bar{x} \leq \hat{q}\hat{x}$. To show it, we have to analyze the effect of increasing q on the monopoly optimum condition with respect to x . Multiplying equation (2) by x , we get

$$x \frac{\partial C}{\partial x} = xqf(xq) \left[1 + \frac{1}{\eta} \right]. \quad (16)$$

Assume that the regulation does not change the volume of services:

$$\bar{x}\bar{q} = \hat{x}\hat{q}.$$

Then we have

$$\hat{x}\hat{q}f(\hat{x}\hat{q}) \left[1 + \frac{1}{\eta} \right] = \bar{x}\bar{q}f(\bar{x}\bar{q}) \left[1 + \frac{1}{\eta} \right].$$

Also if we let

$$\frac{\partial C}{\partial x} = \text{constant (given } q),$$

then

$$x \frac{\partial C}{\partial x} = \text{total costs.}$$

Since for any given amount xq of services produced, the total monopoly costs before regulation were at the minimum, therefore (for $\bar{x}\bar{q} = \hat{x}\hat{q}$):

$$\bar{x} \frac{\partial C}{\partial \bar{x}} \geq \hat{x} \frac{\partial C}{\partial \hat{x}}. \quad (17)$$

The equality in equation (17) exists if for the production of a given amount of services xq different combinations of x and q will cost the same as at the optimum. In this case the preregulation solution for the monopoly was determined for the total amount of service xq but the specific mix of x and q was indeterminate. In this case, both the monopoly and the public are indifferent to regulation. The inequality of (17) contradicts equation (16). Now, let the total services, produced after the regulation, be greater than before the regulation, or $\bar{x}\bar{q} = a\hat{x}\hat{q}$ where $a > 1$. Since we have a declining demand curve, if we assume a declining marginal revenue we find:

$$\bar{x}\bar{q}f(\bar{x}\bar{q}) \left[1 + \frac{1}{\eta} \right] < a\hat{x}\hat{q}f(\hat{x}\hat{q}) \left[1 + \frac{1}{\eta} \right]. \quad (18)$$

However, by the rationale of equation (16), we obtain

$$\bar{x} \frac{\partial C}{\partial \bar{x}} \geq a\hat{x} \frac{\partial C}{\partial \hat{x}}. \quad (19)$$

Once again, (18) contradicts (19). The only solution for $\bar{x}$ in which the marginal condition for x exists under the constraint $\bar{q} > \hat{q}$ is $\bar{x}\bar{q} < \hat{x}\hat{q}$. And, of course, $\bar{x}\bar{q} = \hat{x}\hat{q}$ in the specific case in which everyone is indifferent to regulation (not counting the cost of the regulation itself).

Replace $\frac{\partial C}{\partial x} = \text{constant}$, which was discussed before, by $\frac{\partial C}{\partial x}$

which is an increasing function of x (given q). Note that $\frac{\partial C}{\partial x}$ is the additional cost of increasing x by one unit, namely increasing the number of units of service by q . Therefore, as long as we exclude the corner solutions mentioned before, the marginal cost schedule of producing one unit of service will be higher (or, at the extreme, equal) after regulation than before it. Namely, for any given quantity of services xq we obtain

$$\frac{\partial C(\bar{x}, \bar{q})}{\partial \bar{x}} \frac{1}{\bar{q}} > \frac{\partial C(\hat{x}, \hat{q})}{\partial \hat{x}} \frac{1}{\hat{q}}.$$

This is a more extreme case (from the public's point of view) than the previous one of $\frac{\partial C}{\partial x} = \text{constant}$. Hence, we can apply the same analysis. In short, we shall obtain the conclusion arrived at before that $\bar{x}\bar{q} < \hat{x}\hat{q}$. That is, the total amount of services supplied by a monopoly under constraints on x or on q will be lower than in the absence of regulations.

The above analysis can be restated by the following geometrical presentation: the industry is confronted with a demand curve for services, where one unit of service is one x of $q = 1$. This demand curve is the maximum units of service people want to buy at given prices. Let us denote this demand curve by D and its derived marginal curve by MR . The monopoly has an increasing marginal cost curve (in the extreme case it is constant). For any amount of services xq this curve gives the lowest cost combination (on the margin) of x and q . Regulation sets a restriction on q . Therefore, the post-regula-

tion MC curve will be above the preregulation MC (or at most coincide with the old MC). Regulation affects neither D nor MR . Hence, the post-regulation monopoly optimum at $MR = MC$ will be at a smaller amount of services xq .

■ Another type of quality improvement is that which increases the demand for the product. As an example, take color versus black-and-white T.V. This does not affect the durability, and the quality is not a substitute for the quantity, as in the cases previously discussed.

Let P_q be the price of one unit of product with quality q . Then the demand equation is

$$P_q = g(x, q),$$

with the two conditions

$$\frac{\partial g(x, q)}{\partial x} < 0 \text{ and } \frac{\partial g(x, q)}{\partial q} > 0.$$

As before, the monopoly firm will maximize $[g(x, q)x - C(x, q)]$ with respect to x and q . The two marginal conditions obtained are

$$\begin{aligned} \frac{\partial C(x, q)}{\partial x} &= \frac{\partial g(x, q)}{\partial x} + g(x, q) \\ &= g(x, q) \left[1 + \frac{\frac{\partial g(x, q)}{\partial x}}{g(x, q)} x \right] = P_q \left(1 + \frac{1}{\eta} \right) \end{aligned} \quad (20)$$

and

$$\frac{\partial C(x, q)}{\partial q} = \frac{\partial g(x, q)}{\partial q} x = \frac{\partial P_q}{\partial q} x. \quad (21)$$

Note that the right-hand side of (21) is the increment in revenue the monopoly derives from raising the quality q . It is the increase in the price of a unit of goods times the number of goods it sells. Here there is no adverse effect from the quality improvement.

The two marginal conditions of a competitive market are to equalize the marginal cost with respect to x and the price of a unit of x , given the quality, and to increase q as long as the marginal cost per unit of x with respect to q is lower than the increase in price due to the improved q . The first condition is given by

$$\frac{\partial C(x, q)}{\partial x} = g(x, q) = P_q, \quad (22)$$

and the second condition is given by

$$\frac{\partial C(x, q)}{\partial q} = \frac{\partial P}{\partial q} x. \quad (23)$$

Note the similarity between the monopoly marginal condition for q , given in (21), and the competitive market solution given in (23). However, the monopoly will produce a smaller quantity x than the competitive industry (for the same q), as can be seen by comparing

3. Quality which increases the demand for services

equations (20) and (22), determined by the degree of monopoly $\frac{1}{\eta}$. This, in return, may affect q . The cost function may be one of the following two:⁴

$$\frac{\partial C/x}{\partial q} \text{ is independent of } x, \text{ or} \quad (i)$$

$$\frac{\partial C/x}{\partial q} \text{ is also a function of } x. \quad (ii)$$

(i) *The marginal cost with respect to q , per unit of x , is independent of the level of x .*

This condition can be stated as

$$\begin{aligned} \frac{1}{x} \frac{\partial C(x, q)}{\partial q} &= \theta(q), \text{ or} \\ \frac{\partial C(x, q)}{\partial q} &= x\theta(q), \end{aligned}$$

so that the cost function is of the form:

$$C(x, q) = x\psi(q) + f(x).$$

An example of such a cost condition is a product which has to go through a separate process to improve its quality.⁵

In this type of cost function, the marginal cost with respect to q per unit of x , is independent of the marginal cost with respect to x . Equation (21), for monopoly, is the same equation (23), for competition. The monopolized industry will produce goods of the same quality q as the competitive industry.⁶

(ii) *The marginal cost with respect to q , per unit of x , depends on the level of x .*

Since we are interested in quality, we shall only discuss the effect of this interdependence on the quality q and not on the quantity x .

Let the marginal cost per unit of x , with respect to q , be a decreasing function of x , namely⁷

$$\frac{\partial C(x, q)/x}{\partial x \partial q} < 0.$$

⁴ Throughout, we assume $\partial C/\partial q > 0$.

⁵ A numerical example for cost function of the above type, for which the marginal cost with respect to x and the marginal cost with respect to q are both increasing, is given in the following table:

x/q	1	2	3
1	1	2	4
2	3	5	9
3	6	9	15.

⁶ Note that this optimum q differs from the solution which will be obtained for durability with the same characteristic cost function. Changes in N , durability, affect future sales and costs. Extending the durability for a monopoly means saving the cost of producing future goods. Therefore, the behavior of the marginal cost with respect to x affects durability, even with separation of the marginal costs.

⁷ An example of such a cost function is whenever there are economies of scale

Using (21) or (23), i.e. $\frac{\partial P}{\partial q} - \partial \frac{C(x, q)}{x} / \partial q = 0$, we can by implicit differentiation find $\frac{dq}{dx}$:

$$\frac{dq}{dx} = \frac{\frac{\partial^2 C(x, q)}{\partial x \partial q} - \frac{\partial^2 P}{\partial x \partial q}}{-\frac{\partial}{\partial q} \left[\frac{\partial P}{\partial q} - \frac{\partial C(x, q)}{x \partial q} \right]} \quad (24)$$

By the second order condition for maximization the denominator is positive, and the sign of $\frac{dq}{dx}$ is the sign of the numerator. In the case in which $\partial^2 \frac{C(x, q)}{x} / \partial x \partial q < 0$, to get definite sign we need $\frac{\partial^2 P}{\partial q \partial x} \geq 0$. This is the case where improved quality makes the demand function more elastic. Since $x_m < x_c$ and in this case $\frac{dq}{dx} > 0$, the monopolized industry will fall short of the competitive counterpart not only in x but also in q . Notice that this includes the case where $\frac{\partial P}{\partial q}$ is independent of x . In the case where $\frac{\partial^2 P}{\partial x \partial q} < 0$, that is the case where increased quality increases the absolute value of the slope of the demand function, the result is not unambiguous.

Following similar analysis it is quite clear that if $\frac{\partial^2 C(x, q)/x}{\partial x \partial q} > 0$ and $\frac{\partial^2 P}{\partial x \partial q} < 0$, we shall find $\frac{dq}{dx} < 0$, and $x_m < x_c$ implies $q_m > q_c$. This usually means that if a competitive industry is organized and operated as a monopoly it will produce less x but will raise q . The case $\frac{\partial^2 P}{\partial x \partial q} > 0$ is ambiguous.

□ **Implications for policy.** If the quality improvement increases the whole demand curve of the industry, the following policy comments may be derived from the above analysis:

(1) If $\frac{\partial C/x}{\partial q}$ is independent of x , then, the quality q , is not affected by the market structure, and the quantity x produced. Hence $q_m = q_c$, and there is no place for regulating the quality.

(2) If $\frac{\partial C/x}{\partial q}$ is also a function of x , then from the social point of

in the process of improving quality. It might be that there is a specific process for improving q by each unit. For example, improving q from 1 to 2 is done in process A , and from 2 to 3 in process B , and so on. In each of these processes, economies of scale may exist, although this cost function may be higher for process B than for A , and for process C than for B , etc. In this last example, the marginal cost (per unit x) with respect to q , increases while being a decreasing function of x .

view, the monopoly solution may be distorted in two dimensions: quantity and quality.

Regulating the quantity produced may eliminate the distortion: namely if the monopoly is forced to increase the quantity up to $x_m = x_e$, the marginal cost per unit of x , with respect to q , will change so that the new solution will wind up with increasing quality in the one case, and decreasing quality in the other. On the other hand, setting standards for quality such that $q_m = q_e$ will eliminate the quality distortion, but because the monopoly is confronted with decreasing demand, the quantity x will still fall short of the social optimum. This does not mean that the government should regulate quantity rather than quality. Another relevant factor is the cost of regulating each of these variables.

4. Durability of goods

■ Durability is a type of quality improvement which has the following two characteristics: (1) increasing the durability raises the demand functions for x , or, in other words, the demand is a positive function of the durability N ; and (2) the durability of the product does not affect the supply of the product's services during 1 to $N - 1$ periods, but only the stock the firm is confronted with at period N . Increasing the durability from $N - 1$ to N increases the stock of goods available at period N , and, therefore, might affect the output (and costs) the firm is going to produce at that specific period.

Let x be the quantity of the goods produced in every period of time. It is assumed that the service is proportional to the quantity of this good and with no loss of generalization we choose units such that one unit of goods supplies one unit of service per unit of time. Let P be the price per unit of service, then the demand is given by $P = g(Q)$, where Q denotes the stock of goods available at the given period. If production is spread evenly throughout time and goods are aged according to one horse shay method,⁸ then in a steady state $Q = N \cdot x$. With a rate of interest r the price of a good with a durability of N periods is

$$P \int_0^N e^{-rt} dt = P \frac{1 - e^{-rN}}{r}.$$

Let us analyze the marginal conditions with respect to x and to N for a monopoly firm in a steady state.⁹ As explained, the firm produces x units per period and the available stock is always $Q = N \cdot x$. Now suppose the firm decides to increase production only at period $t = 0$, by one unit of good. Periodic production of subsequent periods does not change from the steady state output of x . As a result of this change, the available stock will be higher by one unit of good in

⁸ Swan [5] has shown the similarity of results as far as durability is concerned—obtained with exponential decay model.

⁹ It is assumed that goods produced have the same durability. It is conceivable that during the path to the steady state—the period in which the stock is accumulated—the best policy of firms is to have other N for goods produced in different periods of time. Therefore, the results presented in the following analysis are not proved for the accumulation period (positive or negative ones) in which N varies. Whenever N is the same in the accumulation period as in the steady state period, since the following analysis of the N refers to any steady state reached, it is also valid for the accumulation period.

each one of the periods from $t = 0$ up to $t = N$. The marginal revenue of the monopolized firm can obtain in each period is $P\left(1 + \frac{1}{\eta}\right)$, where η is the demand elasticity of services per unit of time, and the present value at $t = 0$ of a stream of such marginal revenues obtained in N periods is $P\left(1 + \frac{1}{\eta}\right)\left(\frac{1 - e^{-rN}}{r}\right)$. Therefore, the monopolized firm's marginal cost with respect to x is

$$\frac{\partial C(x, N)}{\partial x} = P\left(1 + \frac{1}{\eta}\right)\left(\frac{1 - e^{-rN}}{r}\right), \quad (25)$$

where $C(x, N)$ is cost function.

If, on the other side, starting from the above described steady state, the firm decides to increase durability N only of goods produced at period $t = 0$, the available stock at the market will not be changed from the steady state level of Q until period $t = N$, but then it will be higher by x , only for one period of time, and then will return to the previous level of Q . The marginal revenue the firm will obtain at that exceptional period is x (namely, the additional available stock) times $P\left(1 + \frac{1}{\eta}\right)$. The present value at $t = 0$ of this amount is $xP\left(1 + \frac{1}{\eta}\right)e^{-rN}$. Hence the monopolized firm's marginal cost with respect to N is:

$$\frac{\partial C(x, N)}{\partial N} = xP\left(1 + \frac{1}{\eta}\right)e^{-rN}. \quad (26)$$

Dividing equation (26) by (25), we find that a monopolized firm has the following optimum condition:

$$\frac{1}{x} \frac{\partial C(x, N)}{\partial N} \bigg/ \frac{\partial C(x, N)}{\partial x} = \frac{re^{-rN}}{1 - e^{-rN}}. \quad (27)$$

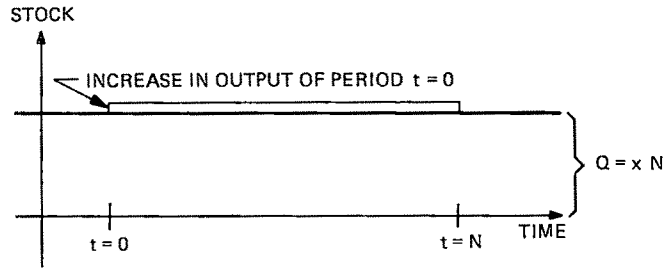
Figure 1 describes the above two optimum conditions.

A competitive firm on the other hand takes the service price or rental price as given. In competitive equilibrium the firm considers its marginal cost per unit of goods with durability N and its price and does not consider its effects on the existing stock of goods available in the future periods. Hence, the competitive firm will equate its marginal cost with respect to x to the price of a unit of x , which is the present value of the rental prices up to N . Secondly, in competitive equilibrium, the firm will increase the durability N as long as the marginal cost per unit of goods with respect to N is lower than the price increment for products with durability N rather than $N - 1$. Therefore, the equilibrium conditions for a competitive firm are, with respect to quantity,

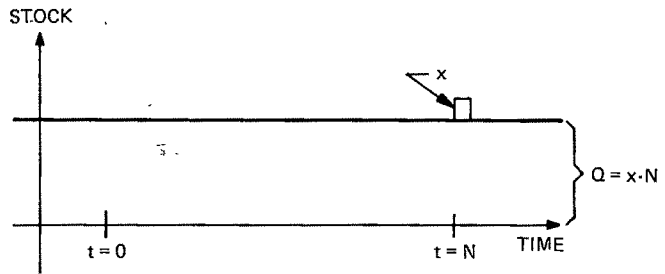
$$\frac{\partial C}{\partial x} = P \int_0^N e^{-rt} dt = P \left[\frac{1 - e^{-rN}}{r} \right], \quad (28)$$

FIGURE 1

- (a) MARGINAL CHANGE IN x – THE NUMBER OF UNITS OF GOODS PRODUCED AT PERIOD $t=0$ IS INCREASED BY 1



- (b) MARGINAL CHANGE IN N – DURABILITY OF OUTPUT OF PERIOD $t=0$ IS EXTENDED BY A UNIT OF TIME



and with respect to durability,

$$\frac{\partial \frac{C(x, N)}{x}}{\partial N} = \frac{1}{x} \frac{\partial C}{\partial N} = \frac{d}{dN} P_N = \frac{d}{dN} \left[P \left(\frac{1 - e^{-rN}}{r} \right) \right] = P e^{-rN}. \quad (29)$$

Dividing equation (29) by (28) we obtain the competitive firm's optimum conditions:

$$\frac{1}{x} \left[\frac{\partial C(x, N)}{\partial N} \right] / \frac{\partial C(x, N)}{\partial x} = \frac{r e^{-rN}}{1 - e^{-rN}}. \quad (30)$$

Notice the similarity between the optimum conditions of a monopolized firm given by (27) and for a competitive firm given by (30), and also between these equations and the optimum ones for quality obtained in previous sections.

Let us compare the equilibrium N that a monopoly and a competitive industry will arrive at under different cost functions.¹⁰

(1) Constant returns to scale cost function, or $C(x, N) = xC(N)$:

Both equations (27) and (30) yield the equation

$$\frac{C'(N)}{C(N)} = \frac{r e^{-rN}}{1 - e^{-rN}},$$

¹⁰ Throughout we assume $\frac{\partial C}{\partial N} > 0$.

and we therefore derive the conclusion that for constant returns to scale $Nm = Nc$, (where subscript c denotes competition, and m denotes monopoly).

(2) Decreasing returns to scale, or $\frac{\partial^2 C}{\partial x^2} > 0$:

(a) Let the marginal cost with respect to N per unit of x be independent of the level of x , namely, $C(x, N) = x\psi(N) + f(x)$.¹¹

In this case $\frac{1}{x} \frac{\partial C}{\partial N} = \Psi'(N)$ is independent of x . So equation (26) for the determination of durability in the monopoly case is $\Psi'(N_m) = p_m \left(1 + \frac{1}{\eta}\right) e^{-rN}$ and in the competitive case is $\Psi'(N_c) = p_c e^{-rN}$.

In our case the marginal cost is rising $\left(\frac{\partial^2 C}{\partial x^2} > 0\right)$, and if the marginal revenue is declining, as it is frequently assumed, then $p_c > p_m \left(1 + \frac{1}{\eta}\right)$ (i.e. the MR in monopoly is lower than the competitive price). Since $\Psi'(N) > 0$, this implies that $N_m < N_c$. On the other hand if the marginal cost is declining (i.e., $\frac{\partial^2 C}{\partial x^2} < 0$), $p_m \left(1 + \frac{1}{\eta}\right) > p_c$, and then $N_m > N_c$. The case of constant marginal cost, (i.e. $C(x, N) = [\Psi(N) + a]x$) yields $N_m = N_c$.

(b) Let the marginal cost with respect to N , per unit of x , depend on the level of x . Here we have a whole variety of results of effects of market structure on N (and x).

Both in the competitive and the monopoly case we have:

$$\frac{\frac{\partial C(x, N)}{\partial x}}{\partial N} - \frac{r}{e^{rN} - 1} \frac{\partial C(x, N)}{\partial x} = 0. \quad (27) \text{ and } (30)$$

Using this equation for the determination of N for each x , we can differentiate and find:

$$\frac{dN}{dx} = - \frac{\frac{\frac{\partial^2 C}{\partial x \partial N}}{\frac{\partial C(x, N)}{\partial x}} - \frac{r}{e^{rN} - 1} \frac{\frac{\partial^2 C(x, N)}{\partial x^2}}{\frac{\partial C(x, N)}{\partial x}}}{\frac{\frac{\partial^2 C(x, N)}{\partial N^2}}{\frac{\partial C(x, N)}{\partial x}} + \frac{r^2 e^{rN}}{(e^{rN} - 1)^2} \frac{\partial C(x, N)}{\partial x} - \frac{r}{e^{rN} - 1} \frac{\partial^2 C(x, N)}{\partial x \partial N}}.$$

If $\frac{\partial^2 C(x, N)}{\partial x \partial N} < 0$ implying that $\frac{\partial^2 C(x, N)}{\partial x \partial N} / \frac{\partial C(x, N)}{\partial x} < 0$ and we

¹¹ Since we are interested in durability, we shall concentrate on marginal cost per unit of x , with respect to N as a function of x , rather than the effect of N on marginal cost with respect to x .

have both $\frac{\partial^2 C}{\partial x^2}$ and $\frac{\partial^2 C}{\partial N^2}$ positive, we find that $\frac{dN}{dx} > 0$. Since $x_m < x_c$ this implies that $N_m < N_c$.

In other cases we do not get the unambiguous result, and it is quite easy to construct cases in which $N_m > N_c$.

Thus, using equation (26) for the determination of N_m for a given x , we can differentiate and find

$$\frac{dN}{dx} = \frac{e^{-rN} \frac{\partial P \left(1 + \frac{1}{\eta} \right)}{\partial x} - \frac{\partial \frac{C}{x}}{\partial x \partial N}}{-\frac{\partial}{\partial N} \left[p \left(1 + \frac{1}{\eta} \right) e^{-rN} - \frac{\partial \frac{C(x, N)}{x}}{\partial N} \right]}.$$

By the second order condition the denominator is positive, and assuming a declining marginal revenue schedule, we find in the case $\frac{\partial \frac{C}{x}}{\partial x \partial N} > 0$ that $\frac{dN}{dx} < 0$. In a similar way we can find that also in the competitive market $\frac{dN}{dx} < 0$. Since $x_m < x_c$, equation (27) implies that $N_m > N_c$. This case, $N_m > N_c$ might in particular occur in cases where $\frac{\partial^2 C}{\partial x^2} < 0$, that is, when we have increasing returns.

References

1. DORFMAN, R. AND STEINER, P. O. "Optimal Advertising and Optimal Quality." *The American Economic Review*, Vol. 44, No. 5 (December 1954), pp. 826-836.
2. KLEIMAN, E. AND OPHIR, T. "The Durability of Durable Goods." *Review of Economic Studies*, Vol. 33, No. 2 (April 1966), pp. 165-178.
3. LEVHARI, D. AND SRINIVASAN, T. N. "Durability of Consumption Goods: Competition versus Monopoly." *The American Economic Review*, Vol. 59, No. 1 (March 1969), pp. 102-107.
4. SCHMALENSEE, R. "Regulation and the Durability of Goods." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 1 (Spring 1970), pp. 54-64.
5. SWAN, P. L. "The Durability of Goods and Regulation of Monopoly." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 347-357.

SHORT ARTICLES, COMMENTS AND REVIEWS

- 250 The Martin Report and its aftermath
M. MENDELSON
- 270 A dynamic model of the regulated firm with a price adjustment mechanism
E. G. DAVIS
- 283 Resource allocation and the regulated firm: comment
A. B. ATKINSON AND L. WAVERMAN
- 288 Resource allocation and the regulated firm: comment on the comments
E. E. BAILEY
- 293 The profit-performance effects of the separation of ownership from control in large U.S. industrial corporations
J. PALMER
- 304 A simple model of firm behavior under regulation and uncertainty
S. C. MYERS
- 316 Pricing of pollution: the Coase Theorem in the long run
H. E. FRECH III
- 320 Pricing of pollution: reply
R. A. TYBOUT
- 322 Regulated utilities and equipment manufacturers' conspiracies in the electrical power industry
E. D. EMERY
- 338 Depreciation in a simple regulatory model
B. L. JAFFEE
- 343 Television's second chance: a retrospective look at the Sloan Cable Commission
H. J. LEVIN

The Martin Report and its aftermath

Morris Mendelson

Associate Professor of Finance
Wharton School
University of Pennsylvania

The confluence of the institutionalization of the equity market and of the development of computer technology is having a drastic organizational impact on the American stock market. The Institutional Investor Study called for a competitive central market system. The Martin Report accelerated the development of such a system, but it is unlikely that the system that will evolve will bear much resemblance to the one recommended in that Report. The fixed commission rate system seems to have little chance of surviving. Full-fledged institutional membership is almost inevitable. The central market system is on the horizon. That might not have been true if the Martin Report had not been written. The Report was to be the solution to the market's ills. Unfortunately, the Report diagnosed healthy competition as a cancerous growth. Fortunately, the error of the diagnosis was recognized, and there is now hope that for the first time in history, the hallmark of American capitalism, the American stock market, may be truly competitive.

1. Introduction

■ In an unintended way the *Martin Report*¹ has been an influential document. It rescued from threatened obscurity the *Institutional Investor Study*.² In spite of a wealth of findings fascinating to the economist, the *IIS* provided little guidance with respect to possible regulations that might alleviate the problems that have arisen out of the institutionalization of the equity markets.³ For years the industry and its regulators had looked to regulation to achieve its objectives. Competition was rarely considered an adequate alternative or even a complimentary force. The *IIS* looked principally to competition to correct the ills that had beset the industry. There consequently seems

AUTHOR'S NOTE: Since the above was submitted (early December 1972), the pace of development has been rapid. The SEC adopted Rule 19b-2 requiring that the exchanges adopt rules that would permit institutions to be admitted to exchanges and required that Exchange members do no more than 20 percent of their brokerage business with affiliated accounts (narrowly defined). This was promptly challenged in court by the PBW Exchange. Banks have begun to wonder whether the Glass-Steagall Act prohibits them from engaging in the brokerage business and, if not, whether failure to open brokerage affiliates might subject them to suits. Representative Moss has introduced H. R. 5050 into the House. The Bill provides, among other things, for competitive commission rates on all transactions after February 1, 1975 (subject to postponement for one year if the SEC finds such postponement in the public interest). The Williams Committee has published a new report and the SEC has released the Policy Statement of the Securities and Exchange Commission on the Structure of a Central Market System. The latter reflects the fear of tampering with a system that works. Looked at from the point of view of the Exchange, the central market system represents a substantial erosion of the Exchange's position. Looked at from the logic of available technology, the central market system involves a host of complications

to prevail a feeling that the study was of limited usefulness. The industry awaited, with baited breath, the *Martin Report*. The thrust of that *Report* was to go back to the good old days.⁴ To all except a segment of the securities industry, the *Report* was a considerable disappointment. It had failed to grasp the realities of the market place today. It failed to recognize the dual impact of the development of communications technology and the institutionalization of the equity market. It did, however, emphasize once again the desirability of a central market system⁵ and this threw into prominence the central theme of the *IIS*, the need for "a strong central market system . . . to which all investors have access, in which *all* qualified broker-dealers and existing market institutions may participate in accordance with their respective capabilities, and which is controlled not only by appropriate regulation *but also by the forces of competition.*"⁶ The italicized words distinguish the views of the two reports.

In this paper, I examine the *Martin Report*. However, instead of confining myself to the contents of that *Report* and its merits and demerits, I use this review as a vehicle for discussing some of the major developments and controversies that have marked the securities industry since the appearance of the *Report*.

Since publication we have had a multi-volume study of the securities industry by the House⁷ and another study in the Senate.⁸ Of particular interest in connection with the latter two is the somewhat controversial report of the Senate Committee,⁹ the Williams

that could be eliminated, if the preservation of the Exchange were not an important consideration. The SEC is not ready to let the Exchange wither away until the SEC is sure that competition will work and until it is sure that self-regulation can be preserved without the Exchange.

The Exchange has offered to trade off its fixed commission rate system in exchange for the abolition of the third market. That may prove to have been a tactical error. The development of a composite tape and composite quotation system has been slower than the SEC hoped, but progress has been made. The Exchange has managed to win some major concessions which may help to preserve it as the primary market. This may, however, prove to no avail if the H. R. 5050 provision that the SEC adopt rules to make sure that investors get the best price becomes law.

There are many differences between the SEC's central market system and my own ATS 6, but the basic difference is more in the distance that has been travelled than in the direction taken. (April 13, 1973)

¹ William McChesney Martin, Jr., "The Securities Markets," a report with recommendations submitted to the Board of Governors of the New York Stock Exchange (August 5, 1971).

² *Institutional Investor Study Report of the Securities and Exchange Commission*, 92nd Congress, 1st session, House Document 92-64. Henceforth, this will be referred to as the *IIS*.

³ *Securities Industry Study*, report of the Committee on Banking, Housing, and Urban Affairs, 92nd Congress, 2nd session, 1972 (henceforth identified as *Senate Report*), p. 37.

⁴ See Mendelson [7] and Farrar [2].

⁵ *Martin Report*, p. 5.

⁶ *IIS*, Volume 1, p. xxv. Italics added.

⁷ Subcommittee of Commerce and Finance of The Committee on Interstate and Foreign Commerce, House of Representatives, 92nd Congress, 1st session.

⁸ Subcommittee on Securities of the Committee on Banking, Housing and Urban Affairs. U.S. Senate, 92nd Congress, 1st session.

⁹ It is somewhat controversial because of the manner in which it was issued. See the several individual views appended to the report.

Bill,¹⁰ and especially the report of the House Committee¹¹ which comes closest to endorsing the recommendation of the *IIS*. There is also the SEC's *Market Structure Report*.¹² This in turn resulted in the appointment of three committees to study the structure, regulation, and governance of a central market system, the development of a comprehensive disclosure system, and the problem of block trading. At this writing the last two committees have filed reports.¹³ All of these documents concur on the need for, or assume the existence of, a central market place.¹⁴

The *Martin Report* contains two broad sets of recommendations: those to be implemented by the Exchange itself and those requiring broader implementation. In Martin's own words, the first set of recommendations called for:

- (1) Reorganization of the New York Stock Exchange.
- (2) Continuous review and emphasis on financial soundness of member firms.
- (3) Additional regulation and improvement of the role of the specialists and block positioners.
- (4) Prohibition of institutional membership and prohibition of member firm management of mutual funds.
- (5) Prohibition of crediting commissions against any fee charged on investment advice.
- (6) Better use of modern communication systems.¹⁵

The second set called for:

- (1) Development of a national exchange system providing a national auction market for each listed security.
- (2) Consideration of increased requirements for entry into the securities business by broker dealers.
- (3) Adoption of appropriate segregation requirements with respect to free credit balances.
- (4) Resolution of the differences which result in unequal regulation and the elimination of the third market, preferably through the development of a national exchange system.
- (5) Additional time to be given to the experiment with negotiated rates before any further change is made.
- (6) The enactment of legislation granting antitrust exemption

¹⁰ This is Bill S.4071, 92nd Congress. The bill was actually sponsored by Senators Bennett and Williams. However, it is generally associated with Senator Williams.

¹¹ Henceforth cited as the *Moss Report*.

¹² Statements of the Securities and Exchange Commission on the future structure of the securities market (February 2, 1972).

¹³ The report to the Securities and Exchange Commission by the Advisory Committee on Market Disclosure on a Composite Transaction Reporting System (July 17, 1972). This report appears in the Bureau of National Affairs (BNA), "Securities Regulation and Law Report," (July 1972), Appendix E. Since this report is only 2 pages, references and citations will not identify the page number; "Report to the Securities and Exchange Commission by the Advisory Committee on Block Transactions," BNA, "Securities Regulation and Law Report," (August 23, 1972), Appendix D. This is also only a 2 page report.

¹⁴ See *Market Structure Report*, p. 7 ff.; the *Senate Report*, p. 33; the *Moss Report* letter of transmittal, pp. xiii ff. References to central markets also appear in the two Advisory Committee reports.

¹⁵ P. 25.

to the exchanges coexistent with Securities and Exchange Commission oversight.

(7) A coordinated effort to eliminate the stock certificate.

(8) Development of a "consolidated exchange" tape.¹⁶

To some extent it is not really possible to separate, as neatly as Martin has appeared to, the problems of Exchange and market reform. In some instances Exchange rules are tantamount to market rules. Coupling a rule regarding the Exchange with a call for a uniform regulation leads to a recommendation for a market rule. Thus, instead of dividing Martin's recommendations as he does, I divide them into the following categories:

(1) Those that apply exclusively to the Exchange.

(2) Those that would apply to the securities industry or a portion of it regardless of the way in which trading was organized.

(3) Those directly relevant to trading on a central market place.

Within the first category is the recommendation of restructuring the Board of Governors of the Exchange. Within the second are recommendations relating to the financial soundness of brokerage houses. The recommendations of the third category relate to market organization.

Our interest is with the latter and in dealing with those it is helpful to compare the Martin recommendations, the *Market Structure Report*, and the *Moss Report*, particularly the first and the last, since the *Moss Report* is the latest document to be published.

■ Quite central to the entire debate regarding the restructuring of the market place is the question of a system of minimum commission rates. No study of restructuring is really possible without devoting considerable attention to that question. The Minimum Commission Rate System is the product of a cartel. As long as the MCRS exists, the cartel exists. As long as the cartel exists, there will at best be linked markets instead of a central market.

In the matter of the MCRS the contrasts in these three reports are striking. Whereas the *Martin Report* calls for "additional time to be given to the experiment with negotiated rates before any further change is made,"¹⁷ the SEC proposes a reduction of the breakpoint from \$500,000 to \$300,000 and announces a determination "to move toward the point at which commission rates on all orders of *institutional size* will be, at least in part, subject to *competitive rates*."¹⁸ To the SEC caution is only required in the timing of the change in the breakpoint. The *Moss Report* categorically rejects the whole concept of an MCRS and recommends that "competitively determined rates should apply to all transactions regardless of size, and that a competitive commission rate system should be phased in without excessive delay."¹⁹ It threatens that if steps are not taken to introduce competitive rates on transactions of small investors, the subcommittee

2. The minimum commission rate system (MCRS)

¹⁶ Pp. 25-26.

¹⁷ *Martin Report*, p. 26.

¹⁸ The *Market Structure Report*, p. 33. Italics added.

¹⁹ *Moss Report*, p. XIV.

will introduce legislation to do so.²⁰ The Williams Bill provides for the reduction of the breakpoint to \$100,000.²¹

A second contrast between the views of the *Martin Report* and the other two is the consistent use in the *Martin Report* of the phrase “negotiated rates” and the almost equally consistent use of the phrase “competitive rates” by the others. Had Martin used the latter phrase, he would have had to eliminate some of the sophistry in his argument. Thus it is irrelevant to the essence of the problem that “for the millions of investors, there will be no negotiation.”²²

In spite of his reluctance to abandon the MCRS, nowhere does Martin address himself to the almost insurmountable problem of rate setting in a fiduciary industry. The difficulties, expressed in the literature, of rate setting in capital intensive industries are discouraging even in that context. In the brokerage industry there seems to be little hope of equitable regulatory rate making.

The SEC’s attitude toward commission rates is not faultless. In examining the totality of the effects of an MCRS, it is concerned with market fragmentation and institutional devices to avoid paying a price in excess of the cost for an execution. Nowhere does the SEC seem concerned with the misallocation of resources resulting from the shelter given inefficient firms.

The SEC clearly displays a bias toward bundling. It notes that “the major thrust of broker-dealer reform should be toward upgrading standards of service to the public, including the provision of adequate information, advice, care, and responsibility. Any changes in the commission structure should not reverse this process.”²³ This philosophy is questionable.

The MCRS has been defended by the Exchange on a host of grounds—grounds which shifted as their weaknesses were exposed. It is impossible to review and evaluate the defense in detail here. The case for and against the MCRS has been discussed extensively.²⁴ The most that can be done here is to note the major points that have been made in favor of the MCRS and to observe that every one of them has been rejected by one or more of its critics. With each point I have associated a footnote indicating where a refutation of the point can be found. Since Friend and Blume²⁵ have the most comprehensive

²⁰ *Ibid.*, p. XV.

²¹ Specifically the Williams Bill provides for unlimited access to Exchange membership for institutions subject to the constraint that members must not trade for affiliated (broadly defined) accounts. It suspends the latter constraint for those who were members of national exchanges prior to October 9, 1972, until the negotiated rate breakpoint is lowered to \$100,000. The *Moss Report* makes a similar proviso except that the proviso applies to all members regardless of the date they joined a national exchange.

²² *Martin Report*, p. 18.

²³ *Market Structure Report*, p. 31.

²⁴ See NYSE, “Economic Effects of Negotiated Rates on the Brokerage Industry, The Market for Corporate Securities, and the Investing Public,” August 1968 and NYSE, “The Economics of the Minimum Commission Rate,” May 1969; Baxter [1]; Mann “A Critique of the NYSE’s Report on the Economic Effects of Negotiated Commission Rates on the Brokerage Industry, The Market for Corporate Securities and The Investing Public,” presented to the SEC (Autumn 1970); Richard B. West and Seha M. Tinc [8]; Mendelson [6]; Stock Exchange Commission Rates, Hearings before the Subcommittee on Securities of the Committee on Banking, Housing and Urban Affairs, U. S. Senate, 92nd Congress, 2nd Session, 1972; and Friend and Blume [4].

²⁵ In [4].

treatment of the subject, all citations to counter-arguments are to their study. In the case of most positions I have added a few observations that I do not believe are duplicative of those made in the works cited.

(1) *The NYSE makes much of the argument that competitive rates would lead to destructive competition.*²⁶

One of the most exasperating aspects of the *Martin Report* is the total lack of substantiating evidence. In spite of the damaging criticism (that the brokerage industry is a declining cost industry which would be subjected to destructive competition if rates were competitive) to which the Exchange case has been exposed, Martin continues to raise the spectre of an industry in which only giants survive.²⁷

In view of the rapid transformation the securities industry is experiencing, one may wonder whether the fall-out that Friend and Blume anticipate might not be a bit on the optimistic side.

A number of developments have reduced the profitability of brokerage activity. Specifically these developments include NASDAQ, new rules with respect to "free capital," new net capital rules, and sharply rising clerical costs.

The advent of NASDAQ has resulted in a much more effective dissemination of information than was hitherto possible. In doing so it has eliminated a great deal of profitable arbitrage that obtained in the pre-NASDAQ days.

Similarly the restrictions in the case of free capital have reduced the profitability of brokerage operations. Free capital has two principal components: (1) the proceeds from lending marginal securities and (2) free credit balances—the funds collected on behalf of customers but not yet disbursed. The latter are especially important and have been the source of roughly two-thirds of the total capital of NYSE member firms that carry public customer accounts.²⁸

Under the new rules and especially proposed Rule 15c3-3 of the 1934 Act, the purposes for which such free capital can be used have been significantly curtailed. The net effect has been to reduce the most attractive form of leverage in American industry, interest-free funds.

The depressing effect of these on brokerage profits has been accentuated by the most stringent and capital rules to which member firms have been subjected. Whereas, formerly the indebtedness ratio could rise as high as 20, the new limit is 15 and several constraints are imposed if the ratio rises to 10. To avoid these constraints most member firms try to keep these ratios closer to 8 than to 10. The result is a further deterioration of the profitability of brokerage houses.

While all of this may result in a more depressing effect on the number of surviving brokerage houses than Friend and Blume anticipated,²⁹ there is no reason to expect the increased fall-out of

²⁶ *Ibid.*, [4], pp. 5–57.

²⁷ *Martin Report*, p. 19. It is only fair to note that only the Baxter and Mann critiques predate the *Report*.

²⁸ SEC "Study of Unsafe and Unsound Practices of Brokers and Dealers," H.R. Doc. No. 92–231, 92nd Congress, 1st Session, pp. 50 ff.

²⁹ Actually this only appears to be the case. The Friend and Blume [4] study

firms noted to be distributed among the various sizes differently from that already anticipated by Friend and Blume, namely randomly.

At this meeting the industry is undergoing considerable trauma. The volume of trading is relatively low. Many firms appear to be financially weak and approaching the danger point. Many are withdrawing or are considering withdrawing from the industry. Many of the older members of the securities industry find the climate uncongenial and would like to withdraw. Wall Street is said to have a lot of "tired" money. The fall-out will be accentuated by competitive rates. There is nevertheless not sufficient reason to believe that the fall-out will be so great as to reduce the number of firms remaining to a level at which the industry can no longer be characterized as competitive.

(2) *Competitive rates would lead to the institutionalization of the market.*³⁰

The contention that competitive rates would significantly accelerate the institutionalization of the market must be premised on the proposition that competitive rates would so *increase* the cost of trading for individuals that it would drive them out of the market. That is not a wholly meaningful proposition. Trades differ widely in character, and the execution cost will depend upon the precise nature of the trade. A trade on margin is different from a cash trade. A limit order does not cost the same to execute as a market order. A trade in which the buyer requires a transfer of title does not cost the same as one in which he is content to leave the stock in street name. If the mix of trades remained unchanged from day to day, a brokerage house could probably determine the average execution cost for trades of a given size and in a vague sort of way what the break-even size of a trade would be. The fact of the matter is that for any such break-even point, the competitive price will be below the current price for some trades of that size and above for others. The following should be read with the above limitations kept in mind.

Friend and Blume felt that the break point between higher and lower rates, consequent to the introduction of competition rates, would be somewhere between \$500 and \$2000.³¹ Once again, it is not clear to what extent Friend and Blume fully considered the recent developments noted above, though possibly the larger figure reflected some of them. Even with the higher breakpoint, the percentage of the market affected may still not be great. In 1970 the odd-lot volume (sales plus purchases) was barely more than 7 percent of the total stock volume.³² Public individuals' trading, however, accounted for 28.4 percent of the total NYSE volume.³³ Thus there were roughly \$3 of round lot trading for every \$1 of odd-lot trading. A round lot trade consisting solely of 100 shares would have been in excess of \$3000.³⁴ This leaves considerable margin for error.

does not explicitly take these factors into account. From this I have inferred that these factors were not a consideration when they speculated on the fall-out of firms.

³⁰ Friend and Blume [4], pp. 65-69; 113-127.

³¹ *Ibid.*, pp. 64-65.

³² NYSE 1971 *Fact Book*, p. 5.

³³ *Ibid.*, p. 52.

³⁴ The average price on the NYSE in 1970 was \$32.10 (*ibid.*, p. 5). If that is

Offsetting this, however, is something Friend and Blume do not mention. It is alleged that one of the distinguishing features of the American market is that stocks are sold, not bought, i.e., trading is the result of encouragement to do so by the account executive or the customer's man. Competitive rates may very well substantially erode the profitability of execution for the registered representative. There could be a substantial reduction in the registered representative's efforts, thus leading to a parametric shift in the demand curve of investors that could outweigh the movement along the demand curve. On the other hand if the price of an execution on small trades did rise to cover the costs, there would be less cause for the firms to discourage their salesmen from handling small orders.

All these changes could take place. There is no concrete evidence they will. We can be sure, however, that (1) if present break-even points reflect inefficiencies that the MCRS permitted to creep into the system, competitive rates will drive the break-even point down, (2) some firms will offer as many bare-bone executions as possible and this will drive rates down, and (3) even if institutionalization were significantly accelerated, the consequences, as Friend and Blume point out, are hardly so forbidding as supporters of fixed rates would have us believe.³⁵

(3) *Competitive rates will lead to a deterioration of auxiliary services.*³⁶

The greatest fear relates to research. In fact, there is no evidence that a very large part of the brokerage community's output has significantly contributed to the information gathering and dissemination process. More important, there is no evidence that in the free competition for the funds represented by the demand for research, the brokerage houses will win out. Brokerage output has been voluminous, but the firms have operated in a captive market. Much of this output has been acceptable because the investor had no choice. In spite of this, private research organizations have flourished. Often brokerage research consists of regurgitating what other organizations have found. The *Special Study* reported that at times even highly reputed brokerage houses simply regurgitated what they had found in the releases of other brokers.³⁷ Many small firms have simply been vehicles for the research output of their correspondents.

Furthermore, a real problem relates to the output of brokerage houses. There is a bias in that output. The real patrons of the brokerage research departments are the customers' men, the security salesmen, not the investors.³⁸ The thrust of the research effort is to provide the sales force with the material with which they can generate trading—which is a negative factor, though rarely recognized as such in the operational efficiency of the market. Given the SEC's attitude towards stability and a wide-spread reluctance on the part of the ordinary investor to go short, a negative report results in a market of

adjusted for the change in the NYSE corporate index the average price ran to \$42.52 and a round lot would constitute a trade in excess of \$4,000.

³⁵ Friend and Blume [4] make point (1) in footnote 1, p. 65 and summarize their arguments regarding point (3) on p. 135.

³⁶ *Ibid.*, pp. 73–74.

³⁷ *Report of Special Study of Securities Markets of the Securities and Exchange Commission*, 88th Congress, 1st Session, House Document No. 95, Part 1, pp. 351–358, and especially p. 353.

³⁸ *Ibid.*, p. 351.

limited potential. Only a fraction of those customers of the house with a given stock in their portfolio can be induced to sell. Furthermore, only a portion of the house's customers will have the stock in their portfolio. Compared with this, a positive report not only enables the customer's man to approach every customer of his with the appearance of doing the investor a favor, but also gives him something with which to approach a new potential customer. The bias against negative reports is accentuated when the company in question is a corporate client of the brokerage house. This is an unscientific observation that would be difficult to prove. Since brokerage houses rarely produce negative reports, it is difficult to show that the fraction of such negative reports on corporate clients was too small to be due to chance.

A large fraction of brokerage output has been devoted to so-called technical analysis. In view of the findings of those who have examined the random-walk nature of stock price changes, we may be very dubious of the value of such technical output and the harm security markets will suffer by contraction of that output.

More important, one may wonder whether much of this research is really desirable. Consider what happens when a large wire house produces a buy recommendation. It makes little difference whether the recommendation contains anything really new (and in view of the disclosure laws and regulations, that is highly unlikely); every sales person in the firm starts urging his customers to buy. This usually creates a temporary trading imbalance and prices will presumably respond to the pressure. The higher prices can be sustained for some time. Whenever the price falls back from a peak, the salesman has the excuse to call his more reluctant customers again to tell them that this is an opportunity to make up for an opportunity missed and that he should get in before the price recovers. This creates wholly unnecessary price movements.

The defenders of MCRS seem to have so little confidence in the value of the research they do that they seem to fear that no one will pay for it and that the research will not be done. This view depends upon the supposition that the marginal cost of the dissemination of research is close to zero, and that the cost could not be recovered in a competitive market. This is empirically improbable, since, as we noted, there now exists a fair number of profitable research publications serving the financial community. Undoubtedly, some of the research now being done by brokers will lose its market because the investor will be free to expend his research funds on the best research he can find, not on the research that happens to be done by his broker. It must not, however, be assumed that all research costs will disappear from the commission rate system. As the SEC's *Market Structure Report* noted, the suitability rules will require brokerage houses to continue to do research in the absence of which they could not determine suitability.³⁹

³⁹ *Market Structure Report*, p. 38. The SEC is thus truly on dubious grounds in defending bundling. The quality of broker services is not likely to deteriorate. Care in execution is something that can be expected as long as the investor is paying for execution. Indeed, when practically all the broker is delivering is execution, he must be that much more careful. When an investor is buying an auxiliary service and paying with executions, he may tolerate sloppier executions than he would otherwise.

The pure custodial function should soon become obsolete. There is no excuse

(4) *Competitive rates will result in a loss of scale economies.*⁴⁰

Friend and Blume already note the significance of the "current and prospective developments in an expanded NASDAQ" to the advantages the NYSE has historically had in providing good markets.⁴¹

The advantages would be further eroded if the *Moss Report* recommendation that the specialists' books be broadcast to all market-makers is implemented.

If, furthermore, brokers made their entries into electronic books directly, the costs involved in maintaining exchanges could be eliminated. Since most market-makers in the system envisioned by Moss would make markets in many more stocks than the typical specialists, each market-maker could trade a smaller fraction of the total in each stock and still maintain sufficient volume to realize most of the potential economies. To postulate that in such a system all trades must be funneled through a single market-maker is to postulate that not only do units costs of trading fall monotonically with increasing volume, but they do so without an asymptote within the available volume of trading. That is not plausible. More than that, the deterioration of advantages produced by the kind of technological change under consideration has little to do with competitive rates. It is a development that is likely to occur even if the MRCS is sustained.

(5) *Competitive rates would impair the efficiency of the securities markets.*⁴²

(6) *Competitive rates would impair the primary market.*⁴³

In this connection and in connection with the question of institutional membership, our attention is often drawn to the European capital markets as an example of what should be avoided. There is some misapprehension about the ability of European capital markets to absorb equity issues. If we consider the principal EEC countries, we find that from 1965 to 1969 equity issues constituted 17.9 percent of the domestic market security issues, whereas the corresponding figure for the United States was only 13.5 percent.⁴⁴ To be sure, the comparability of these figures leaves much to be desired. They are,

for customer securities' remaining much longer in the physical possession of the broker. The custodial function should be transferred to a central depository as soon as possible. This would result in a much more efficient settlement in clearing process. Even if for some transitory period, the custodial function remained with the brokerage houses rates can be unbundled. The custodial function is not transferable. The house can therefore make a separate charge. Indeed, the broker may find it cheaper to remain custodian than to transfer the share. The only apparent difficulty in the matter of unbundling is in the matter of research. It might be argued that since the output is transferable, the brokerage house cannot capture its true value. This line of argument overlooks the value of the timeliness of the research. Second and third hand analyses of investments in closely watched stocks are not worth much. Even allowing for the deficiencies of the model sometimes used in examining the spread of the impact of news on prices, it is quite clear that the prices fairly rapidly reflect news. Thus while a report is transferable, timeliness is not.

⁴⁰ Friend and Blume [4], pp. 80 ff.

⁴¹ *Ibid.*

⁴² *Ibid.*, pp. 76-106; 381-386, 122-127.

⁴³ *Ibid.*, pp. 107-113; 136 ff.

⁴⁴ OECD *Financial Statistics*, No. 4 (Dec. 1971), p. 514.

however, suggestive that the fears of developing into a European type capital market are exaggerated on that score at least. It should be noted that if the country where institutions have least influence on the market mechanism—the Netherlands—is omitted, equity issues rise to 20+ percent of domestic issues.⁴⁵

There is thus little ground for believing that competitive rates will have all the frightful consequences alleged. However, competitive rates will reduce some of the conflict of interest that plagues the securities industry and will clear the air for some important market reforms.

The debate raging around the MCRS seems to be shifting against the NYSE. In the end, the decisive factor in the debate may not be the force of the logic, but whether or not the arguments are persuasive to those with political clout.

The Friend and Blume study is the most comprehensive and best study to date but is still not definitive. A truly definitive case would be extremely difficult to produce. As good as the Friend-Blume study is, it can be no better than the data on which it is built. Many of the data are from the income and expenditure accounts the brokerage houses filed with the SEC. Establishment of a reasonably uniform method of accounting in the brokerage industry is still a thorny problem. The data must be considered suspect to some extent at least. There are, however, no known biases in the data that would lead one to expect the Friend-Blume results to have been different, if greater data comparability had been achieved.

Although the case against the MCRS has been weakened by recent developments, that weakening does not appear to be anywhere near fatal. In considering the case for and against the MCRS, it should not be forgotten that the Exchange is asking for a shelter from competition. As a number of protagonists have argued, "since the Exchange is asking for some very special privileges, the burden of proof must surely lie with the Exchange not the critics."⁴⁶

3. A central market system

■ Martin has given a high order of priority to the development of a central market system. He specified that such a system "provide a *single* national auction market for each security qualified for listing."⁴⁷ In presenting the case he failed to clarify the essence of an auction market and to defend its superiority over other arrangements.⁴⁸ He called for subordination of broker-dealers, specialists, and other professionals, to the interest of the public. Yet, many of his recommendations clearly subordinate the public interest to the interest of the Exchange community. He called for "fair" commission rates, not competitive commission rates. He did, however, enunciate the need for "a market which provides for equal access by all investors to material information about both the market itself and the securities traded in that market, including centralized reporting of price and volume of all trades."⁴⁹

⁴⁵ *Ibid.*

⁴⁶ See Mendelson [6], pp. 94-95.

⁴⁷ *Martin Report*, p. 5.

⁴⁸ For clarification of the issue see Mendelson [6], pp. 12-14; *Senate Report*, pp. 34-37; and statement of Feuerstein [3], pp. 2970-2977.

⁴⁹ *Martin Report*, p. 5.

His concept of integration was peculiar indeed. In calling for a single market in each listed security he was, in effect, recommending the end of competition with NYSE-listed securities from the third market and from the regional exchanges. He did not recommend the absorption into the Exchange of the market-making power represented by the over-the-counter firms or the regionals. In preserving the specialists system, with or without competing specialists, he was recommending the preservation of a franchise system and artificially constraining the number of market makers. It is hardly surprising that Martin's central market system suggestions were not well received by either the regional exchanges or the third market firms.⁵⁰

Martin's protective attitude towards the specialists is in striking contrast with that found in a report of the SEC's advisory committee on block trading. The latter are willing to let block positioners register as "upstairs market-makers." In effect this would permit them to engage in off-the-floor trading by quoting on NASDAQ. They would thus go into competition with the specialists. Specialists in turn are to be permitted to become upstairs market-makers for securities other than those in which they specialize. In trading of these stocks NYSE Rule #113 would not apply.⁵¹ These recommendations presume the implementation of a composite tape reflecting all trades.

Martin seems to have had a clearer concept of the essence of a central market than either the SEC or its advisory committee. He seemed to have recognized the desirability of a confluence of all bids and asks as well as a focal point for information on executions. Though the advisory committee recommendations seemed to strengthen significantly the competitive character of market making, the proposed structure leaves the displacement process partial and necessitates the unnecessary interpositioning of market-makers.⁵² The committee is not really to be faulted for recommending a structure in which there is absolutely no possibility of confluence of orders. What they have recommended is indeed an improvement. They did, however, fail to recognize the shortcomings of a central market system as they visualize it. They do not assume, as they seem to believe, "the existence of a central securities market involving competition within a market rather than competition between markets." The competition is indeed between markets. A composite tape merely results in more closely linking the markets.⁵³ However, as long as

⁵⁰ It is interesting how a spokesman for the NYSE, which has defended the MCRS in terms of avoiding hardship for existing firms, should so blandly call for the destruction of this important segment of the market.

⁵¹ The effect of this rule is to prohibit specialists from trading directly with institutions.

⁵² NYSE Rule #394 is a bit of an anomaly. The rule was promulgated to confine trading in listed stocks to the floor of the Exchange. It is an anticompetitive rule and fortunately ineffective. The result has been fragmentation. It could only result in complete centralization at a high cost. See the discussion of that rule in Mendelson [6], pp. 29-30.

⁵³ See SEC 1934 Act Release No. 9850, November 8, 1972 re proposed Rule 17-a-15. The call for a consolidated tape actually predates the *Martin Report*. However, it would seem that the *Martin Report's* recommendations accelerated the development of this rule. Interestingly enough the *Martin Report's* recommendation is ambiguous. Since it called for confining the trades of each security to a single market, the consolidated tape would be an innocuous device at best reducing the number of instruments one had to watch. The Rule 17-a-15 requires

displacement is only partial, the markets have not been integrated into a single entity. In a way this is evidenced by the recommendation of the SEC's Advisory Committee on Market Disclosure on a Composite Transaction Reporting System that the market place in which transactions take place be identified.

The NYSE persists in its attempts to destroy or hamper the third market. In its reply to the SEC's proposed Rule 17-a-15 relating to a composite tape, it endorsed "the concept of a system of competing quotations, *but one which is centered in the exchange market.*"⁵⁴ This position is defended on the grounds that inclusion of third market-makers "would lead to the demise of the auction market system."⁵⁵ Unfortunately, while the Exchange may represent one of the greatest auction markets the world has ever seen, its spokesmen have often failed to appreciate the essence of an auction-market and display an astonishing lack of faith in the vitality and efficiency of the institution they so vigorously defend. They just do not seem to appreciate the fact that the essence of an auction-market is that it provides for the displacement of tenders by superior ones.⁵⁶ If they did, their spokesmen would not talk about a "quotation" network for dealers acting as principals in NYSE-listed stock which would "eliminate the auction market at least on one side of each trade."⁵⁷ Nor would they talk about the substitution of a dealer-market system for the present auction exchange. Denying third market-maker's access to the network reduces the opportunity for displacement.

4. The specialist

■ All of the aforementioned reports attempt to perpetuate and even strengthen the specialist system. Unfortunately, the specialist system under the existing regime is not well suited to block trading,⁵⁸ and under a regime such as the one envisaged by the *Moss Report* its existence seems to be threatened.⁵⁹

The growth of institutional trading has added a new dimension to the securities market. Institutions want to trade in large blocks. Why is not our concern. The fact is they do. While some specialists have demonstrated their talent and willingness to handle blocks, many have not. Specialists operate under three handicaps in the treatment of blocks. The NYSE Rule #113 precludes direct access to major parties interested in the stock in question. This rule is partially designed to reduce the conflict of interest resulting from the specialist's acting as both agent and principal. It is also designed to protect the income of brokerage houses which must consequently act as intermediaries between the institutional traders and the specialist. This gives rise to a second, though less important, difficulty spe-

a trade in a security to be reported on the consolidated tape regardless of the market in which it took place.

⁵⁴ Memorandum to NYSE Board of Directors submitted to SEC September 25, 1952, italics added.

⁵⁵ *Ibid.*

⁵⁶ Feurerstein [3] and my own adaption of his argument in [6], pp. 12-13.

⁵⁷ Memorandum, *op. cit.* Actually nearly half of all shares now traded in the NYSE involve a principal on at least one side of the market. See New York Stock Exchange 1972 *Fact Book*, pp. 74-75.

⁵⁸ Mendelson, [6], pp. 25-31 and *Senate Report*, pp. 40 ff.

⁵⁹ The relevant provisions of the *Moss Report* are discussed below. The merits of this remark will then become apparent.

cialists face in the institutionalized market. Sometimes these intervening brokers botch an execution and claim it is the fault of the specialist. Traders then become reluctant at times to allow the broker to bring the block to the floor of the Exchange.

Because the specialist has no direct access to the institutional market, he has no sales force. Therein lies the advantage of the block house. It does have an established network of communications with institutions. The sales force, being in daily contact with these institutions, has a fairly good idea of which institutions are interested in which stocks. If the specialist is indeed an efficient market-maker, an exchange of information with a block positioner can be helpful to both. Cooperative efforts seem to be relatively rare, but when they have taken place, the results seem to have been highly satisfactory.

The specialist, but not the block trader, operates under an obligation to maintain continuity in the market. The advantage of the specialist over the block trader is that he is the focal point for the flow of orders.

Block trading is a quasi-underwriting activity. The conventional negotiations lead to a guarantee by the assembler, the block positioner approached by a seller in a block trade, of a minimum price to the institution in question.⁶⁰ The assembler then tries to assemble buyers and stands ready to absorb any part of the block that cannot be disposed of.

The interest of the buyer arises, at least in part, from the possibility of acquiring a substantial amount of stock at what is, in effect, a discount from market. Kraus and Stoll have shown that blocks traded roughly one commission below earlier executions and that price recovery is usually fairly rapid.⁶¹ There is then an opportunity for some instant profit.⁶² The block positioner is presumably compensated for his risk by subsequently laying off at a recovered or partially recovered price the position he has built up.

The constant worry of the block positioner is an over-hang. The appearance of a new block after he has executed a block has a further depressing effect on price (and on his position) and generates ill will on the part of those who participated on the passive side of the first block. Furthermore, if institutional traders suspect that another block will be forthcoming, they will tend to withhold their purchases in order to see the price effect of the block that, in effect, "cleans up the street," thus making it additionally difficult for the block positioner.

It is a widely accepted dogma of the Street that it is easier to execute a block of 50,000 shares than two blocks of 25,000 each in the same stock. A specialist acting solo because of Rule # 113 finds it more difficult to know whether or not there is an over-hang. Because of these shortcomings, specialists generally cannot handle block trading without the supplementary efforts of the block positioner.

At the same time, block positioners can create problems for the specialist. It is often profitable for the specialist to participate in

⁶⁰ Most blocks seem to be initiated on the selling side. It will be convenient to discuss this case.

⁶¹ Kraus and Stoll [5].

⁶² A profit, incidentally, which must by the logic of the system be lost when the stock is subsequently disposed of in a block sale.

block trades. This participation may not be desirable from the point of view of the block positioner. Having arranged to dispose of an entire block, he may find that he has to give some of it up to the specialist. To avoid this, and for other reasons, the assembler often brings a trade to a regional exchange for execution. In the absence of a consolidated tape, news of such trades and of the execution price does not immediately reach the specialist. Such activities can, in Martin's words, "frustrate public participation in the market and the orderly operation of the market."⁶³ Martin offered no specific remedies for the difficulties and instead called for "rules and regulations specifying the qualifications for those firms which may act as block positioners and defining their obligations to the public market place . . . , and a closer working relationship between block positioners and specialists."⁶⁴ Again there is a reaffirmation of the role of the specialist.

Many of the difficulties alluded to above would be eliminated if the recommendations of the Advisory Committee on Block Trading and the *Moss Report* were put into effect.

Briefly, the relevant recommendations of the Advisory Committee are:

- (1) Possible modification of NYSE Rules # 113, 394 (for off-the-floor executions by member firms), and 438 (for members' advertising bids and offers for over-the-counter markets), and consideration of amendments to Section 11(b) of the Securities Exchange Act of 1934.
- (2) Permission for block positioners to register as upstairs market-makers, and to advertise on NASDAQ.
- (3) Permission for specialists to register as upstairs market-makers in stocks in which they do not specialize.
- (4) Maintaining the specialist as the focal point for reporting transactions and as the interface between large and small transactions. Registered upstairs market-makers would be required to probe the market on the floor allowing the pre-existing public orders to displace the block transactions.

The *Moss Report* differs in some respects. It goes further. Among other things it substitutes the concept of the member of the system for seats on the Exchange and proposes access to the central market for all qualified broker-dealers. It recommends that "neither the specialist nor other market-makers would be required to make a market for any security against their will,"⁶⁵ and that the "specialist's book would be computerized and made available to other specialists and market-makers" and that there would be a requirement that no order be executed within the system until better bids and offers in the book were satisfied.⁶⁶

It should be noted that there is a subtle difference in the recommendation of the Advisory Committee with respect to NYSE Rule # 113. The Advisory Committee would require a revision of the rule to permit the specialist to deal with institutions in stocks in which they are upstairs market-makers and only "explore directly the size and timing of blocks coming into the market." The *Moss Report* recom-

⁶³ *Martin Report*, p. 14.

⁶⁴ *Ibid.*

⁶⁵ *Moss Report*, p. XIV.

⁶⁶ *Ibid.*

mends its removal. The difference, however, is not likely to be important because the specialists system is not likely to survive. Once the upstairs market-maker is allowed, commission rates become competitive, and the book is made available to all makers, then the specialist franchise seems to have no basic source of value. This is especially true if the NYSE specialist's book has to be consolidated with the books of the specialists on the regionals before broadcast. The investor can establish his priority by placing his bid or ask with a regional specialist just as easily as with the New York Stock Exchange specialist. Given the abandonment of Rule 394, when the investor approaches a broker with stock to sell, it is incumbent upon the broker to check all markets. Orders cease to flow almost automatically to the NYSE floor.⁶⁷ Specialists will find that if brokers are to place orders through them, they must be competitive with each other and with upstairs market-makers. Their charges must be competitive. However, if the specialist agency function is solely to monitor the flow of limit orders and broadcast them, he can easily be replaced by a computer into which the orders are placed directly by the brokers.⁶⁸

The removal of that function from the specialist would make him essentially no different from any upstairs market-maker. Although the *Moss Report* hedged, it really foresaw the end of the exchanges. It observed that "although . . . exchanges would not in themselves be declared obsolete or merged with one another, we visualize that, eventually membership in an exchange would be superseded by membership in a central market system. To the extent that trading is not automated and physical 'floors' are necessary, the concept of a 'seat' on the Exchange would be replaced by the right to execute transactions on the floor and to charge other members of the system a reasonable amount in 'floor brokerage' for doing so."⁶⁹ It is not clear even in the absence of automated trading that a floor is necessary. Automating the input to the book really eliminates the need for a physical floor.

One must, however, never underestimate the resilience of men whose capital is threatened. We must therefore consider the possibility that the specialist will find some way to sustain the value of the franchises to execute agency transactions. Some specialists may feel that they can capitalize on their marketing ability best by maintaining informal relationships with a number of block positioners. In the stocks assigned to these specialists, competitive market-making will be sustained. However, the specialist may find that the liaison they build up with the block positioner is sustainable only so long as they do not build up an upstairs trading force of their own, and this may discourage them from going it alone. The remaining specialists have the choice of building up an upstairs force and not attempting to establish any special liaison or simply merging with some block house. In the latter event the merging of a good specialist

⁶⁷ There often exist potential executions off the book which, if captured, will prevent price fluctuations that might otherwise have occurred. Provision for such stabilization will have to be built into the mode of executing market orders if the efficiency of the market is to be maximized.

⁶⁸ It should be noted that AMEX's AMCODE is designed so that brokers can enter their orders directly. In this matter no judgment is involved and there is no reason for hesitating to put these orders on a computer. The interposition of the specialist can thus be dispensed with.

⁶⁹ *Moss Report*, p. XIV.

with a good block positioner might create an unbeatable combination which would destroy competitive market-making in the stocks assigned to the specialist in question.

5. Institutional membership

■ The question of institutional membership is broader than the question of membership in an Exchange. It is a question of access to the market-maker.

Martin opposed membership on three ostensible grounds: the concentration of economic power which might result, the possible evolution of a dealer rather than an agency market, and the need to preserve the difference between securities businesses and other businesses.⁷⁰ None of these grounds really merits serious consideration. The hollowness of the first is evident in the Exchange's endorsement of the SEC's rule that institutionally affiliated brokers be prohibited from doing more than 20 percent of their business with their parents.⁷¹ This would force the large institutions to merge with the largest brokerage houses with whom they could arrive at an agreement. This would create larger firms than would unrestricted entry.

Strangely enough, Martin failed to consider adequately the possibility of the conflict of interest in the money management-brokerage activity merger. He proposed merely to require member firms to divest themselves of mutual fund operations. This would have left them with pension funds which are economically more significant than mutual funds.

He brushed aside the question of the MCRS as a consideration. But there is reason to believe that is precisely what concerned the Exchange community.⁷² Since the *Martin Report*, there has been considerable pressure to reduce the breakpoint for negotiated trading to \$100,000. With that the concern shifts to retaining, for the Exchange community, the business of institutions. Hence, the 80-20 rule of the SEC. If the conflict of interest consideration were really dominant, the SEC should have ruled as the *Moss Report* recommended and as Senator Williams' bill provides: unlimited institutional access to the market, with all direct dealings between brokerage houses and affiliated accounts (fairly broadly defined) prohibited.⁷³ More important, if the conflict of interest problem was truly a dominant consideration, the separation of agency and principal transaction should have been considered seriously. There the potential for conflicting interest is far greater.⁷⁴ Even the *Moss Report* ignores this problem. The *Moss Report* rationale for its position is seen in the proviso that the prohibition of trading with affiliated accounts "should not take effect until the breakpoint for competitively determined rates has reached the \$100,000 level."⁷⁵

The Williams Bill, however, provides for money managers' paying in excess of competitive rates in exchange for research, providing full

⁷⁰ *Martin Report*, pp. 16-17.

⁷¹ Release 9716, August 2, 1972, regarding Proposed Rule 19b-2.

⁷² If that were indeed not the case, it is difficult to understand why Martin would single out the prohibition of crediting commissions against management fees.

⁷³ *Moss Report*, pp. XVI. In the SEC proposed rule 19b-2, affiliated accounts are more narrowly defined than in the *Moss Report* and the Williams Bill.

⁷⁴ See Mendelson [7], p. 71.

⁷⁵ *Moss Report*, p. XIV. The Williams Bill also makes this proposal.

disclosure is made. This is an unfortunate provision. It involves a tied sale. The managers should be free to buy the best research and executions available. Only incidentally will they be supplied by the same firm. The alleged difficulty with hard dollar payments for research is that these dollars would be paid by the management company, whereas soft dollars are paid for by the beneficiaries of the managed funds. The feared consequence is that the management company would buy less research and the beneficiaries would suffer. If this is a genuine danger, surely it can be avoided by the simple process of permitting the management company to pass on to the fund research costs incurred.

The solution of the *Moss Report* and the Williams Bill is clearly superior to that of the SEC. The conflict of interests that permeates money management brokerage affiliation is more carefully avoided, the affiliated accounts are assured competitive rates, institutions are not restricted in the diversification activity, and a potentially large source of capital is made available to the securities industry.

There are two potential flaws in this view. One is that the solution leaves open the possibility of back scratching by affiliated brokers. The other is the assumption of a competitive market. Above I discuss the possibility of specialists-block positioners mergers destroying some competition. Under these circumstances, the prohibition of trading with affiliated accounts could make some of the trading unduly expensive. However, this argument was based on the development of some unforeseen tenacity on the part of the specialist. Given the nature of the operating rules of the market, that probability should be discounted.

It remains to be seen which of the recommendations will be implemented and whether the rules ultimately promulgated are in fact conducive to a competitive market. If they are not, the investing public will again have been injured by the restrictions on trading through affiliated brokers. This trading will be more expensive. Then, we have to face the conflict of interests question squarely. The fact of the matter is that recent studies have not shown that broker-affiliated institutions have had significantly higher turnover ratios than others.⁷⁶ This does not mean that the potential for abuse should not be curbed, but under the circumstances envisaged restrictions on affiliated account trading would not seem to be in the public interest.⁷⁷

■ In dealing with the question of equal regulation one must bear in mind in whose interest the regulation has been formulated. There

6. Equal regulation

⁷⁶ IIS, p. 90. Excessive turnover is not the only potential abuse of trading for affiliated accounts. The broker may give the affiliated accounts unwarranted priorities in trading and the dissemination of research results. Recent investigations have cast less light on this problem than on the turnover problem. Giving unwarranted priorities need not be confined to firms trading for affiliated accounts. The problem may arise whenever a broker's mix of unmanaged accounts varies substantially in size. Then, in order to keep the very large accounts happy, the broker may deal with them preferentially.

⁷⁷ I am, of course, referring to agency-type transactions. I do not question the merits of Section 17 of the Investment Company Act which prohibits principal transactions between broker dealer and an affiliated investment company. Nor do I question the merits of applying the substance of that section to other affiliated accounts.

In contrast Martin recommends that the prohibition not be so extended. It

is often made an arbitrary and unwarranted assumption that complete uniformity is desirable. Of the six items mentioned, the first four are designed to protect the minimum commission rate system and the Exchange community. The last two unquestionably have merit. Indeed, one might further recommend that all short sales rules should be uniform.

Second, the *Martin Report* never really considered the meaning of the concept of equal regulation. It seemed to be taking about identical regulation.⁷⁸ However, identical regulation of nonequals does not really constitute equal regulation. As the *Senate Report* notes, "the call for 'equal regulation' is no guide for regulatory action, unless it is first determined who is equally situated in terms of the purposes of the regulation."⁷⁹ Similarly the composite tape Advisory Committee speaks of appropriately equal regulation.

7. Antitrust laws

■ Martin's suggestion that the Exchange be specifically exempted from the antitrust laws was not particularly well received outside of the Exchange community. It was categorically rejected in the *Moss Report*. The only concession made there is the recommendation that "Congress should consider . . . whether treble damages in antitrust actions challenging exchange rules should be mandatory or discretionary with the district court judge."⁸⁰

8. Conclusion

■ There are several areas not touched by this paper. These include recommendations for the reorganization of the New York Stock Exchange, recommendations regarding the capital requirements and related matters of brokerage firms, the problems of foreign brokers, problems relating to the clearing system and communications technology in general. About the latter in particular, the *Martin Report* had regrettably little to say.

In this paper, I have confined myself to those recommendations closely related to the development of an efficient central market system. Two things are remarkable about the chain of events since the *Report* was released. It is remarkable how little of what was recommended proved to be acceptable. Nevertheless, it is remarkable how far we have moved in the direction of implementing a competitive central market place. It is significant that these innovations are seriously being considered by law makers, rule makers, and industry leaders. Even when the concept of a competitive central market place was discussed in the *IIS*, it had an academic tinge. The discussion has moved to the political plane and that is a prerequisite for the implementation of the conditions under which a competitive central market system can develop. It is, however, no guarantee for the implementation of those conditions. The political clout of the Exchange com-

is in this same passage that he recommends that brokers be "prohibited from crediting commissions against any fee charged for investment advice," (p.18). This was a blatant effort to preserve the MCRS. The horror that MCRS defenders express at the idea of institutions' joining exchanges to save money seems to suggest that they consider cost-cutting immoral.

⁷⁸ See *Martin Report*, p. 14, where "unequal" is equated with "different."

⁷⁹ *Senate Report*, p. 43.

⁸⁰ *Moss Report*, p. XVII.

munity must not be discounted. The *Moss Report* is only a report as yet, and consideration of the Williams Bill has been postponed to the next session. It is not yet legislation, and its contents may change radically before it becomes law.

References

1. BAXTER, W. F. "NYSE Commission Rates: A Private Cartel Goes Public." *Stanford Law Review*, Vol. 22, No. 4 (April 1970), pp. 675-712.
2. FARRAR, D. E. "Wall Street's Proposed Great Leap Backward." *Financial Analysts Journal*, Vol. 27, No. 5 (September-October, 1971), pp. 14-61.
3. FEUERSTEIN, D. M. Statement in *Study of the Securities Industry*. (Hearings before the Subcommittee on Commerce and Finance of the Committee on Interstate and Foreign Commerce.) House of Representatives, 92nd Congress, 2nd Session, pp. 2970-2977.
4. FRIEND, I. AND BLUME, E. "The Consequences of Competitive Commissions on the New York Stock Exchange." Mimeographed, Rodney L. White Center for Financial Research, The Wharton School, 1972.
5. KRAUS, A. AND STOLL, H. R. "Price Impacts on Block Trading on the New York Stock Exchange," *Journal of Finance*, Vol. 27, No. 3 (June 1972), pp. 569-588.
6. MENDELSON, M. "From Automated Quotes to Automated Trading: Restructuring the Stock Market in the U.S." *The Bulletin*, Nos. 80-82 (March 1972), pp. 89-95.
7. ———. "Nostalgia versus the Computer: The Issue of Stock Market Reform." *Wharton Quarterly* (Fall 1971), pp. 32-40.
8. WEST, R. R. AND TINIC, S. M. "Minimum Commission Rates on New York Stock Exchange Transactions." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 2 (Autumn 1971), pp. 577-605.

A dynamic model of the regulated firm with a price adjustment mechanism

E. G. Davis

Carleton University

Instead of the Averch-Johnson assumptions of instantaneous regulation of the rate of return, but no regulation over the setting of individual rates, this note assumes that the regulator applies the new rates to the test year's output to predict the revenue requirements under the allowed rate of return. A mathematical model of the firm is then worked out. When the A-J view of the regulated firm is dynamized in a rather straightforward fashion, the A-J point no longer has any special significance: it is not the stationary point of the dynamic version of the model. While it is true that the stationary-point does exhibit over-capitalization in the A-J sense, the regulatory agency can use the extent to which it adjusts prices as a further control variable to influence the stationary point.

1. Introduction

■ By now, the Averch-Johnson view of the regulated firm¹ is well-known: the firm operates, given the technical facts of life (a production function), to satisfy the wishes of the market (a demand curve) as it attempts to maximize "profits" subject to a rate-of-return constraint imposed by the regulator. In addition, the world is static—all relevant relationships are not only time-invariant but are completely known to all concerned. The only constraint dictated by the regulatory agency is that the firm must earn no more than a predetermined "fair rate of return on capital." The resultant equilibrium position exhibits "economic inefficiency" in the sense that there is a divergence between the ratio of the firm's marginal physical (or revenue) products and the ratio of the (constant) factor prices.

This model gives the firm wide scope for action: subject to the "fair-rate-of-return" constraint, it can "set" prices at whatever level it wishes. In rate-setting, the regulator has no role to play! A more realistic model would surely have the regulator choosing rates explicitly to force the firm to a position where it earns precisely the desired rate of return.

Moreover, a look at the real world would lead any observer to see imperfections in the assumption that the regulator reacts instantaneously and successfully to bring the actual and desired rates of return

This work was done while the author was associated with the Economic Analysis and Graphical Research Department at Bell Telephone Laboratories and the Economics Studies Group at A.T.&T. The views presented here are solely those of the author and in no way express the opinion of Bell Telephone Laboratories or of the American Telephone and Telegraph Company. The author would like to thank the referee for vastly improving this version over an earlier draft.

¹ In [1].

into equality whenever they differ. Obviously a useful generalization of the basic model would be an explicit recognition of the lags inherent in the regulatory process.

At least two such attempts to model this "regulatory lag" have appeared in the literature.² Each paper introduces simply a temporal lag—the regulatory agency does not monitor the firm's actual rate of return continuously, but instead makes the exactly appropriate rate changes at discrete time intervals. Although at each point of time the regulator "knows" the level of rates at which the firm will earn precisely the fair rate of return, the focus of his attention is elsewhere in the interim periods between rate adjustments.

This paper will offer a slightly different view of this regulatory rigidity. In the actual process of rate making a rather crude rule seems often to be followed: roughly, this year's output at next year's prices will result in the firm's earning precisely the desired rate of return on capital. This view seems to be shared by Baumol and Klevorick:

As we see it, regulation proceeds, roughly, by taking the latest information available on company costs and demands, and resetting prices so that the firm's *current* rate of return—based on its current capital stock, technology, demand conditions, and so forth—is reduced to the "fair" rate level.³

(Obviously, in the absence of a perfectly inelastic demand curve, this represents an adjustment mechanism rather than an immediate solution.) This practice may simply be a convenient approximation or may reflect uncertainties about what rate structure will admit of exactly the "fair rate of return on capital's" being earned. To model this process, we assume that the regulator follows a generalization of this rule for rate changes in an effort to "lead" the firm to the appropriate rate of return. However, the regulated firm is assumed to have perfect knowledge about both demand and regulatory behavior to aid in its profit-maximizing decisions.

■ In the spirit of the Averch-Johnson approach the model is formulated as follows:

- (1) A production function relating output (Q) to inputs of capital (K) and labor services (L):

$$Q = F(K, L), \quad (1)$$

where it is assumed that $F_L > 0$, $F_K > 0$, $F_{LL} < 0$, $F_{KK} < 0$, and $F_{KL} = F_{LK} > 0$.

- (2) An adjustment mechanism for the output price (P) dictated by the regulator:

$$\dot{P} = S \left(\frac{rK + wL}{Q} - P \right), \quad (2)$$

where

S = a positive constant (with the interpretation of a speed of adjustment controlled by the regulatory agency),
 r = the "fair rate of return on capital," and
 w = the (constant) wage rate.

2. A model of the regulated firm

² See [2] and [4].

³ In [3], p. 184.

Obviously if $S = 1$, this equation is simply the formalization of the "this year's output at next year's prices, etc." rule of thumb.

(3) A demand function:

$$P = P(Q), \quad P' < 0, \quad (3)$$

where we assume PQ to be concave with respect to K and L :

$$\frac{\partial^2(PQ)}{\partial^2 L} < 0, \quad \frac{\partial^2(PQ)}{\partial^2 K} < 0, \quad \frac{\partial^2(PQ)}{\partial K \partial L} > 0.$$

(4) The firm's objective is to maximize the present value of future profits, π , where

$$\pi \equiv \int_0^\infty e^{-\rho t} (PQ - wL - RK) dt \quad (4)$$

and

ρ = the (constant) discount rate applied to future profits and

R = the firm's "cost of capital."

As is common, we assume $r > R$ so as to allow the firm to earn positive profits.

The firm is always constrained to satisfy demand and is assumed to know all the relevant information about the demand and production relationships and the price adjustment mechanism.

For the moment it is useful to reformulate equations (1) and (3) as:

$$K = G(Q, L), \quad G_1 > 0, \quad G_2 < 0, \quad (1a)$$

where G_i denotes the partial derivative with respect to the i th argument,⁴ and

$$Q = D(P), \quad D' < 0. \quad (3a)$$

The firm's problem is now formulated in terms of an optimal control problem which can be solved by Pontryagin's Maximum Principle.⁵ The regulated firm chooses $L(t)$ so as to maximize π :

$$\max_{L(t)} \int_0^\infty e^{-\rho t} \{PD(P) - wL - RG[D(P), L]\} dt$$

subject to:

$$\dot{P} = S \left\{ \frac{rG[D(P), L] + wL}{D(P)} - P \right\},$$

with given initial conditions, $P(0)$ and $L(0)$.⁶ The Hamiltonian is

$$\begin{aligned} \mathcal{H} \equiv & e^{-\rho t} \{PD(P) - wL - RG[D(P), L]\} \\ & + \gamma S \left\{ \frac{rG[D(P), L] + wL - PD(P)}{D(P)} \right\}, \end{aligned}$$

⁴ This formulation implies that the firm can adjust its capital stock instantaneously and frictionlessly to the desired level. Given the extensive literature supporting the contrary view, the author can only support this approach by virtue of its simplicity and hope that it does not do too much violence to reality.

⁵ In [5].

⁶ It should be noted that $F_K = \frac{1}{G_1}$ and $F_L = -\frac{G_2}{G_1}$.

where γ is a Lagrangian multiplier. For a maximum, the first order conditions must be satisfied. Hence,

$$\frac{\partial \mathcal{L}}{\partial L} = 0 \quad (5)$$

$$\dot{\gamma} = - \frac{\partial \mathcal{L}}{\partial P}. \quad (6)$$

It is useful to transform γ by

$$\lambda \equiv e^{\rho t} \gamma. \quad (7)$$

Now differentiating equation (7) using equation (6) yields

$$\dot{\lambda} = \rho \lambda + e^{\rho t} \dot{\gamma} = \rho \lambda - e^{\rho t} \frac{\partial \mathcal{L}}{\partial P}. \quad (8)$$

When the calculations are performed, we obtain:

$$\frac{\partial \mathcal{L}}{\partial L} = 0 = \frac{w + RG_2}{w + rG_2} - \frac{S\lambda}{D(P)}, \quad w + rG_2 \neq 0. \quad (9)$$

One further transformation simplifies matters:⁷

$$C \equiv \frac{\lambda}{D(P)}. \quad (10)$$

Differentiating yields:

$$\dot{C} = \frac{\dot{\lambda}}{D(P)} - \frac{\lambda D'}{[D(P)]^2} \dot{P}. \quad (11)$$

Equations (2), (8), (9), and (10) can be substituted into equation (11) to yield:

$$S(w + rG_2)D(P)\dot{C} = \rho D(P)(w + RG_2) + (R - r)SD'G_1 \left[w + \frac{D(P) + PD'G_2}{D'} \frac{G_2}{G_1} \right]. \quad (12)$$

The dynamic process, with given initial conditions, is described by equations (2) and (12). Although these expressions are somewhat complicated, some idea of the dynamic behavior around the steady-state point can be obtained graphically and this will be done in the following section.

■ From footnote 6, we know that the marginal product of labor is simply $[-G_2/G_1]$. Hence, the curve which is the locus of points where the wage equals the marginal revenue product of labor is

$$w + \left[\frac{D(P) + PD'G_2}{D'} \frac{G_2}{G_1} \right] = 0. \quad (13)$$

⁷ C has the interpretation of the marginal value to the firm of the price per unit of output divided by the output.

3. A look at the dynamic behavior

We observe that:

$$w + \frac{(D + PD') G_2}{D' G_1} = w - (P + P'Q)F_L \equiv \phi(K, L). \quad (14)$$

From the concavity assumptions of equation (3), we find that

$$\phi_L > 0 \quad \text{and} \quad \phi_K < 0,$$

and hence that

$$\left. \frac{dL}{dK} \right|_{\phi=0} > 0.$$

Above the $[\phi(K, L) = 0]$ curve, in $L - K$ space, the wage exceeds the marginal revenue product of labor. Below, the converse is true.

Furthermore, the $[w + RG_2 = 0]$ curve is the expansion path and we observe that

$$w + RG_2 = w - \frac{RF_L}{F_K} \equiv \Psi(K, L). \quad (15)$$

With our assumptions about the production function, we can calculate that

$$\Psi_L > 0 \quad \text{and} \quad \Psi_K < 0,$$

and hence conclude that $\left. \frac{dL}{dK} \right|_{\Psi=0} > 0$.

We have a similarly shaped curve for

$$w + rG_2 = w - \frac{F_L}{F_K} \equiv \tilde{\Psi}(K, L). \quad (16)$$

In $L - K$ space, Ψ and $\tilde{\Psi}$ are positive above the $[\Psi = 0]$ and $[\tilde{\Psi} = 0]$ curves, respectively, and negative below, and the $[\tilde{\Psi} = 0]$ curve lies above the $[\Psi = 0]$ curve, since $r > R$.

Equation (12) can be rewritten as

$$SQ\tilde{\Psi}(K, L)\dot{C} = \rho Q\Psi(K, L) + (R - r)S\alpha(K, L)\phi(K, L), \quad (17)$$

where $\alpha(K, L) \equiv D'G_1 < 0$.

Equivalently, we obtain:

$$\begin{aligned} \frac{\tilde{\Psi}(K, L)}{(R - r)\rho\Psi(K, L)} \dot{C} \\ = \frac{\rho}{(R - r)S} + \frac{\alpha(K, L)\phi(K, L)}{Q\Psi(K, L)} \equiv \mathcal{E}(K, L). \end{aligned} \quad (18)$$

At the A-J point, the intersection of $[\dot{P} = 0]$ and $[\phi(K, L) = 0]$; $\phi = 0$, $\Psi < 0$, $\tilde{\Psi} > 0$. Therefore $\dot{C} > 0$.

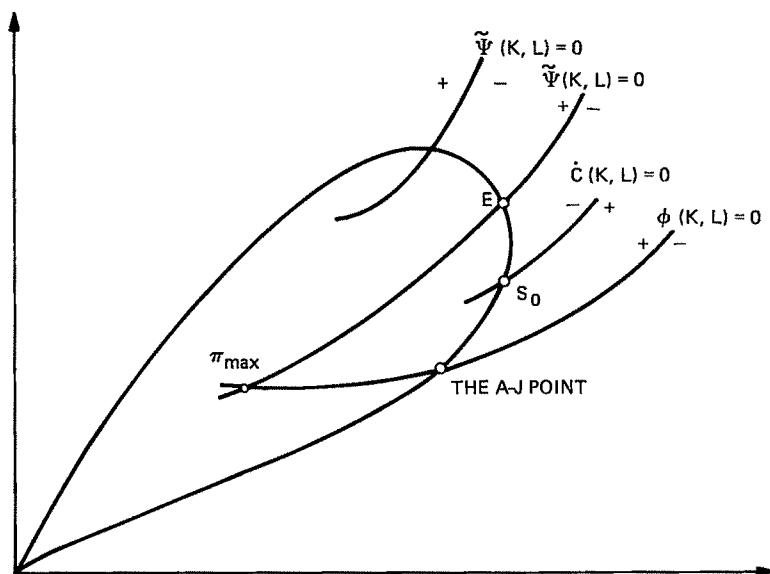
At the efficient point E , the intersection of $[\dot{P} = 0]$ and $[\Psi(K, L) = 0]$, $\phi > 0$, $\Psi = 0$, and $\tilde{\Psi} < 0$. Therefore, $\dot{C} < 0$.

It is shown in Appendix 1 that it is reasonable to take,

$$\mathcal{E}_K < 0, \quad \mathcal{E}_L > 0,$$

and hence $\left. \frac{dL}{dK} \right|_{\mathcal{E}=0} > 0$.

FIGURE 1

THE STATIONARY POINT OF THE MODEL — S_0 

All of the above results are illustrated in Figure 1, following the approach introduced in Zajac's paper.⁸

We can draw three conclusions:

- (1) There exists a stationary point on the constraint curve between E and the A-J point.
- (2) At the stationary point, the marginal revenue product of labor is less than the wage and the marginal revenue product of capital is less than the cost of capital.
- (3) At the stationary point, labor usage is higher than at the A-J point but lower than at E , while output is smaller than that at the efficient point while larger than that at the A-J point.

For the purpose of studying stability of the model as specified in equations (2) and (12), it is obviously convenient to transform back from $L - K$ to $L - P$ space using the transformation:

$$L = L$$

$$K = G[D(P), L].$$

In $L - P$ space we can define

$$\tilde{\mathcal{E}}(P, L) \equiv \mathcal{E}(K, L),$$

and obtain

$$\tilde{\mathcal{E}}_P = \mathcal{E}_K G_1 D' > 0$$

and

$$\tilde{\mathcal{E}}_L = \mathcal{E}_K G_2 + \mathcal{E}_L > 0.$$

Therefore $\left. \frac{dL}{dP} \right|_{\tilde{\mathcal{E}}=0} < 0$. Furthermore, let us define

$$N(K, L) \equiv rK + wL - QP(Q)$$

⁸ In [7].

FIGURE 2

A TRANSFORMATION INTO
L-P SPACE

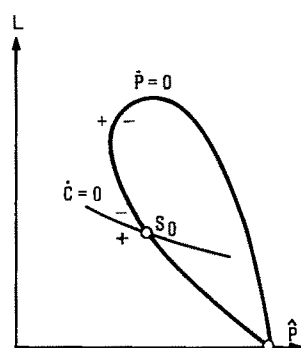
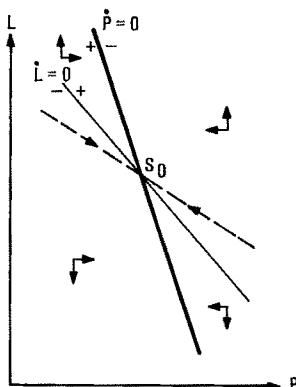


FIGURE 3

STABILITY IN A LOCAL REGION
ABOUT THE STATIONARY POINT



4. A comparative statics result

and

$$\tilde{N}(P, L) \equiv N(K, L).$$

At the stationary point, we can determine that

$$N_L = w - (P + QP')F_L = \phi(K, L) > 0$$

and that

$$N_K = r - (P + QP')F_K > 0.^9$$

These too can be transformed into $L - P$ space and we obtain

$$\tilde{N}_P = N_K G_1 D' < 0$$

and

$$\tilde{N}_L = N_K G_2 + N_L < 0.^{10}$$

These results are illustrated in Figure 2.¹¹

If we differentiate equation (9), we obtain

$$\dot{L} = \frac{(w + rG_2)^2}{w(R - r)G_{22}} S\dot{C} - \frac{G_{21}D'}{G_{22}} \dot{P}. \quad (19)$$

With the generalized Cobb-Douglas production function used in Appendix 1, $G_{21} < 0$ and $G_{22} > 0$, so we see that the coefficients of both $\dot{C}$ and $\dot{P}$ are negative. Looking again at Figure 2, it is apparent that the $[\dot{L} = 0]$ curve lies between the $[\dot{P} = 0]$ curve and the $[\dot{C} = 0]$ curve and thus must have a negative slope.

An appropriate choice for the initial value of L , given an initial value of P in a local region around the stationary point, will allow the stationary point to be reached with motion as shown by the dotted curve in Figure 3.

■ In the previous section it was shown that the stationary point of this model exhibits overcapitalization in the A-J sense. We know that

$$w + RG_2 < 0, \quad (20)$$

which is equivalent to

$$\frac{\frac{\partial Q}{\partial L}}{\frac{\partial Q}{\partial K}} > \frac{w}{R}. \quad (21)$$

Presumably, the regulatory agency is interested in minimizing the extent of this "inefficiency" as well as providing an upward bound on the rate of return on capital.¹² Moreover, in addition to choosing a

⁹ This follows from $N_L > 0$ and $\tilde{\Psi}(K, L) < 0$ at S_0 .

¹⁰ This follows by direct substitution using $\tilde{\Psi}(K, L) < 0$ at S_0 .

¹¹ Since we have assured a straight line demand curve in the Appendix, $\hat{P}$ is simply the minimum price such that $D(P)$ is zero.

¹² It must be noted that the discussion will be only about the properties of the stationary point—and will say nothing about the length of time required to get arbitrarily close to it. Presumably an alteration in the price adjustment mechanism would change both the degree of "inefficiency" at the stationary point and the speed of convergence. Conceivably changes may work in opposite directions—a rule leading to a more efficient stationary point may exhibit slower convergence. These dynamic complications will be ignored but ought to be kept in mind to qualify the conclusions reached.

“fair rate of return,” the regulator has a further control variable, the speed of adjustment constant, S , to affect the ultimate equilibrium.

From equations (2) and (12) we see that this stationary point is defined by

$$PD(P) = rG[D(P), L] - wL = 0 \quad (22)$$

and by

$$\frac{\rho}{S(r - R)} = \frac{(D + PD')G_2 + wG_1D'}{D(w + RG_2)}. \quad (23)$$

In Appendix 2 these two equations are totally differentiated with respect to ρ/S , and it is shown that under rather general conditions

$$\frac{d(G_2)}{d(\rho/S)} > 0. \quad (24)$$

In other words, the smaller the value of S chosen by the regulators, the closer the steady state position will approach the efficient point.¹³ In fact it is not S itself that matters, but rather ρ/S . The less the future matters for the firm (the higher it is discounted), the closer the stationary point will be to the efficient point. This would suggest that if it is the steady-state position which is of importance to the regulator (with the length of time spent in transition less significant), the regulator would do well to follow a fairly passive policy, changing prices less rapidly than would be suggested by assuming a perfectly inelastic demand curve and setting S equal to one.

■ Obviously, the assumptions made in this paper are sufficiently crude that it should not be read for its policy implications. It is, rather, one more example that dynamic analogues to static models can yield different results. When the A-J view of the regulated firm is dynamized in a rather straightforward fashion, the A-J point no longer has any special significance: it is not the stationary point of the dynamic version of the model. While it is true that the stationary point does exhibit overcapitalization in the A-J sense, the regulatory agency can use the extent to which it adjusts prices as a further control variable to influence the stationary point.

5. Summary

Appendix 1

■ In all that follows we will assume, for simplicity, a straight line demand curve.

We have

$$\mathcal{E}(K, L) \equiv \frac{\phi(K, L) \propto (K, L)}{\Psi(K, L)F(K, L)},$$

where $\propto(K, L) \equiv D'G_1 = \frac{1}{P'F_K}$.

¹³ Echoing footnote 7, presumably a smaller S is associated with a slower speed of convergence—meaning that the relatively less inefficient point may be approached more slowly.

This result is rather unsurprising since it is really just a generalization of the well-known result that the perfectly competitive firm (which faces given prices which are beyond its control, i.e., $S = 0$) is efficient.

Differentiating, assuming $P'' = 0$ where appropriate, we obtain:

$$\Psi_L = \frac{-R}{(F_K)^2} [F_K F_{LL} - F_L F_{KL}] > 0,$$

$$\Psi_K = \frac{-R}{(F_K)^2} [F_{LK} F_K - F_L F_{KK}] < 0,$$

$$\phi_L = -(P + P'Q)F_{LL} - 2(F_L)^2 P' > 0,$$

$$\phi_K = -(P + P'Q)F_{LK} - 2F_L F_K P' < 0,^{14}$$

$$\alpha_L = \frac{-F_{KL}}{P'(F_K)^2} > 0,$$

$$\alpha_K = \frac{-F_{KK}}{P'(F_K)^2} < 0.$$

At the stationary point S_0 , we have the following information about signs:

$$\begin{array}{ll} \alpha < 0 & \alpha_L > 0 \\ & \alpha_K < 0 \\ \phi > 0 & \phi_L > 0 \\ & \phi_K < 0 \\ \Psi < 0 & \Psi_L > 0 \\ & \Psi_K < 0. \end{array}$$

Now we shall differentiate $\mathcal{E}(K, L)$:

$$\mathcal{E}_L = \frac{1}{(F\Psi)^2} [F\Psi\phi\alpha_L + F\Psi\alpha\phi_L - \phi\alpha\Psi F_L - \phi\alpha F\Psi_L].$$

Let us first consider the term $(\Psi F_L + F\Psi_L)$:

$$\begin{aligned} \Psi F_L + F\Psi_L &= \left(w - \frac{RF_L}{F_K}\right)F_L - \frac{FR}{(F_K)^2} [F_K F_{LL} - F_L F_{KL}] \\ &= wF_L - \frac{R}{[F_K]^2} [(F_L)^2 F_K + FF_K F_{LL} - FF_L F_{KL}]. \end{aligned}$$

For help in these computations, we assume a production function of generalized Cobb-Douglas form

$$F(K, L) = K^a L^b, \quad 0 < a < 1, \quad 0 < b < 1,$$

$$\text{where } F_K = \frac{aF}{K}, F_L = \frac{bF}{L}, F_{LL} = b(b-1)\frac{F}{L^2}, F_{KK} = a(a-1)\frac{F}{K^2},$$

$$\text{and } F_{KK} = \frac{abF}{KL}. \text{ Hence, we can conclude that } \Psi F_L + F\Psi_L = wF_L$$

$$- \frac{RF_{LL}F}{F_K}.$$

¹⁴ This follows since $\phi(K, L) = w - (P + P'Q)F_L = w - \frac{\partial(PQ)}{\partial L}$ and we have assumed that $\frac{\partial^2(PQ)}{\partial L \partial K} > 0$.

Next we shall consider the term ($\alpha \phi_L + \phi \alpha_L$):

$$\alpha \phi_L + \phi \alpha_L = \frac{1}{P'(F_K)^2} \{F_K[-(P + P'Q)F_{LL} - 2(F_L)^2P'] - F_{KL}[w - (P + P'Q)F_L]\}.$$

Now

$$\begin{aligned} \mathcal{E}_L = & \frac{[-(P + P'Q)F_{LL} - 2(F_L)^2P']}{F\Psi P'F_K} \\ & - \frac{F_{KL}}{F\Psi P'(F_K)^2} [w - (P + P'Q)F_L] - \frac{\phi \alpha}{(F\Psi)^2} wF_L \\ & + \frac{\phi \alpha}{(F\Psi)^2} \frac{RF_{LL}F}{F_K}. \end{aligned}$$

The first and third terms are positive. The second and fourth terms added together are:

$$\begin{aligned} & \frac{\phi \alpha}{(F\Psi)^2} \frac{RF_{LL}F}{F_K} - \frac{F_{KL}}{F\Psi P'(F_K)^2} [w - (P + P'Q)F_L] \\ & = \frac{\phi}{\alpha F\Psi^2 F_K} [RF_{LL} - F_{KL}\Psi] \\ & = \frac{\phi}{\alpha F\Psi^2 F_K} \left[-wF_{KL} + R \left(F_{LL} + \frac{F_{KL}F_L}{F_K} \right) \right]. \end{aligned}$$

Since $\left(F_{LL} + \frac{F_{KL}F_L}{F_K} \right)$ has the sign of $(2b - 1)$, we can only conclude that $\mathcal{E}_L > 0$ if $b < \frac{1}{2}$.¹⁵

Now we consider the other partial derivative:

$$\mathcal{E}_K = \frac{1}{(F\Psi)^2} [F\Psi \alpha \phi_K + F\Psi \phi \alpha_K - \alpha \phi \Psi F_K - \alpha \phi F\Psi_K].$$

The second term in brackets is positive while all the others are negative.

Consider the term $[F\Psi \phi \alpha_K - \alpha \phi F\Psi_K]$:

$$\begin{aligned} & F\phi[\Psi \alpha_K - \alpha \Psi_K] \\ & = \frac{F\phi}{P'(F_K)^3} [(wF_K - RF_L)(-F_{KK}) + RF_{LK}F_K - RF_LF_{KK}] \\ & = \frac{F\phi}{P'(F_K)^3} [RF_{LK}F_K - wF_KF_{KK}] < 0. \end{aligned}$$

So we know unambiguously that $\mathcal{E}_K < 0$.

¹⁵ While this could well be possible in an industry using capital heavily, it is rather unappetizing nonetheless. The author can only plead that it is only a sufficient condition and that possibly, buried in the complex computations, there lurks a weaker and more justifiable condition.

Appendix 2

■ The stationary point, S_0 , is defined by

$$\dot{P} = 0 \quad (\text{A1})$$

and by

$$\dot{C} = 0. \quad (\text{A2})$$

Hence, from equations (2) and (12), we get:

$$rG + wL - PD(P) = 0 \quad (\text{A3})$$

and

$$X \equiv \frac{\rho}{S(r - R)} = \frac{[D(P) + PD']G_2 + wG_1D'}{D(P)[w + RG_2]}. \quad (\text{A4})$$

Now, we differentiate equations (A3) and (A4) by X . We get (assuming $D'' = 0$)

$$\begin{bmatrix} A & B \\ H & J \end{bmatrix} \begin{bmatrix} \frac{dP}{dX} \\ \frac{dL}{dX} \end{bmatrix} = \begin{bmatrix} 0 \\ \{D(P)[w + RG_2]\}^2 \end{bmatrix},$$

where

$$A \equiv rG_1D - (D + PD'),$$

$$B \equiv w + rG_2,$$

$$H \equiv D(w + RG_2)[(D + PD')G_{12}D' + 2G_2D' + wG_{11}(D')^2] \\ - [(D + PD')G_2 + wG_1D'] [w + RG_2 + DRG_{12}]D',$$

and

$$J \equiv D(w + RG_{12})[(D + PD')G_{22} + wD'G_{12}] \\ - [(D + PD')G_2 + wG_1D']DRG_{22}.$$

If we define

$$M \equiv AJ - BH,$$

then we obtain

$$\frac{dP}{dX} = \frac{-(w + rG_{12})D^2(w + RG_{12})^2}{M} \quad (\text{A5})$$

and

$$\frac{dL}{dX} = \frac{D^2(w + RG_2)^2[rG_1D' - (D + PD')]}{M}. \quad (\text{A6})$$

However, we are interested in the sign of $\frac{dG_2}{dX}$:

$$\begin{aligned} \frac{dG_2}{dX} &= G_{12}D' \frac{dP}{dX} + G_{22} \frac{dL}{dX} \\ &= \frac{D^2(w + RG_2)^2}{M} Z, \end{aligned} \quad (\text{A7})$$

where

$$Z \equiv G_{22}[rG_1D' - (D + PD')] - D'G_{12}(w + rG_2). \quad (\text{A8})$$

Using equation (A4) to substitute for $(D + PD')$ in equation

(A8) we find that

$$Z = \frac{D'(w + rG_2)[G_1G_{22} - G_2G_{12}] - XDG_{22}(w + RG_{12})}{G_2}. \quad (A9)$$

For help with signs, we again assume a production function of generalized Cobb–Douglas form:

$$Q = K^a L^b, \text{ where } 0 < a < 1 \text{ and } 0 < b < 1. \quad (A10)$$

It is simple to show that with this production function

$$\begin{aligned} G_1G_{22} - G_2G_{12} &> 0, \\ G_{22} &> 0. \end{aligned} \quad (A11)$$

In addition, we know that at S_0 , the stationary point,

$$\begin{aligned} D' &< 0, \\ w + rG_2 &< 0, \end{aligned}$$

and

$$w + RG_2 < 0.$$

Looking back at (A9) we see that

$$Z < 0,$$

and thus,

$$\operatorname{sgn} \left[\frac{dG_2}{dX} \right] = -\operatorname{sgn} [M]. \quad (A12)$$

If expressions for $(D + PD')$ and $[(D + PD')G_2 + wG_1D']$ from equation (A4), are substituting into M , after tedious but straightforward calculations we find that

$$\begin{aligned} N &\equiv \frac{M}{D(w + RG_2)} = - \left(\frac{XD}{G_2} \right)^2 (w + RG_2)wG_{22} \\ &\quad - XD(w + RG_2) \frac{wD'}{(G_2)^2} [G_{12}G_2 - G_1G_{22}] \\ &\quad + (w + rG_2)D'[Xw + G_2(RX - 2)] \\ &\quad + (w + rG_2)D'DX \frac{w}{(G_2)^2} [G_1G_{22} - G_{12}G_2] \\ &\quad + (w + rG_2) \left(\frac{D'}{G_2} \right)^2 w[G_1G_2G_{12} - G_2^2G_{11} - G_1^2G_{22} + G_2G_1G_{12}]. \end{aligned}$$

Using the information in (A11), we can see that the first, second, and fourth expressions are all positive. Further calculations on the production function in (A10) show that the square bracket in the last expression is negative and hence the whole expression is positive. If we can safely assume that

$$2 > RX, \quad (A13)$$

we can conclude that N is positive or equivalently that M is negative.

Recalling (A12) we would see that

$$\frac{dG_2}{dX} > 0.$$

The question of whether the inequality in (A13) is true can only be settled empirically. However, it seems to be true for reasonable values of the parameters.¹⁶ Assuming this inequality and a production function of generalized Cobb–Douglas form, we have demonstrated that the value of G_2 at the stationary point and X (or ρ/S , given r and R) are positively related.

¹⁶ Taking a simple example ($\rho = 0.1$, $r = 0.08$, $R = 0.07$, and $S = 1$), we see that $RX = 0.7 < 2$.

References

1. AVERCH, H. AND JOHNSON, L. L. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol. 52, No. 5 (December 1962), pp. 1053–1069.
2. BAILEY, E. E. AND COLEMAN, R. D. "The Effects of Lagged Regulation in an Averch-Johnson Model." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 278–292.
3. BAUMOL, W. J. AND KLEVORICK, A. K. "Input Choices and Rate of Return Regulation: An Overview of the Discussion." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 162–190.
4. KLEVORICK, A. K. AND BAUMOL, W. J. "Regulation of Public Utility Profits and Operating Efficiency." Unpublished manuscript, 1968.
5. PONTRYAGIN, L. S., BOLTYANSKII, V. G., GAMBRELIDZE, R. V., AND MISHCHENKO, E. F. "The Mathematical Theory of Optimal Processes." Tr. by K. N. Trilogoff, New York, 1962.
6. TAKAYAMA, A. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol. 59, No. 3 (June 1969), pp. 255–260.
7. ZAJAC, E. E. "A Geometric Treatment of Averch-Johnson's Behavior of the Firm Model." *The American Economic Review*, Vol. 60, No. 1 (March 1970), pp. 117–125.

Resource allocation and the regulated firm: comment

A. B. Atkinson

University of Essex

and

L. Waverman

University of Toronto

This comment presents an analysis of the effect of regulation on the sales-maximizing firm, and shows that the models used in earlier works have been incorrectly specified. It is demonstrated that under relatively weak conditions, regulation via a "fair" rate of return constraint has no influence on the behavior of the firm, and that any tendency to under- or overcapitalization would be present even if there were no regulation.

■ Much of the literature on regulation is based on the assumption that the regulated utility maximizes profits, and does not allow for the possibility that it may pursue other goals such as the maximization of sales, output, or the rate of growth. However, if these latter goals are a more accurate characterization of the behavior of modern corporations than the conventional assumption of profit maximization, as has been argued by Baumol, Marris, Williamson, and others, then it is clearly important to examine the relationship between regulation and alternative objectives. One major exception to the neglect of alternative goals for the firm is the article by Bailey and Malone,¹ who have considered a range of objectives² and shown that the behavior of a regulated utility depends sensitively on the management objectives. In fact they reach the striking conclusion that:

if a firm is constrained to a fair return on invested capital, we have found that the profit-maximizing firm would have an incentive to overcapitalize, while the sales- or output-maximizing firm would have an incentive to undercapitalize.³

Unfortunately, their analysis of the sales maximization case is misleading, since the model is not correctly formulated. The result they describe is only one of a number of possible cases, and under quite reasonable assumptions it ceases to be valid. The aim of this note is to present a correct analysis of the effect of regulation on a sales maximizing firm. The note examines the effect of regulation on the output level and input choices of the firm and brings out the important point that regulation may well make no difference to its behavior—a point which was completely overlooked by Bailey and Malone.⁴

The authors are grateful to A. K. Klevorick for his helpful comments on an earlier draft.

¹ In [2].

² See also Kafoglis [8], Zajac [10], and Baumol and Klevorick [5], Section 7.

³ See [2], p. 141.

⁴ There is a passing reference to this point in Edelson [6], but he does not present a systematic analysis of the sales-maximization case.

1. Introduction

2. Formulation of the model

■ The formulation by Bailey and Malone of the behavior of the regulated sales-maximizing firm is as follows:

$$\max R(Q) \quad (1)$$

$$\text{subject to: } R(Q) - wL \leq sK, \quad (2)$$

where $R(Q)$ denotes the revenue as a function of output (Q). Output, in turn, is a well-behaved function of labor (L) and capital (K), satisfying

$$\begin{aligned} Q(0, K) &= 0, & Q(L, 0) &= 0 \\ Q_L &> 0, & Q_K &> 0. \end{aligned}$$

Labor is purchased from competitive markets at price w , and the cost of capital is denoted by i . The regulatory constraint (2) is assumed to take the form of a permitted "fair" rate of return s on capital and it is assumed that $s > i$.⁵ It is on the basis of this model that they argue that the regulated firm will undercapitalize: i.e., use a capital-labor ratio lower than that which minimizes cost for the output level it chooses to produce. This result is the precise opposite of that found by Averch and Johnson⁶ for the profit-maximizing firm.

This formulation of the problem is, however, an inadequate representation of the firm's behavior. In particular, from Baumol⁷ onwards, the theories of sales maximization have all depicted the firm as maximizing sales *subject to a minimum profit constraint*:

$$\max R(Q) \quad (3)$$

$$\text{subject to: } R(Q) - wL \geq \rho K, \quad (4)$$

where ρ is the required rate of return on capital.⁸ It is their neglect of (4) that led Bailey and Malone astray—they assumed in effect that the unregulated firm maximized R without any constraint on the minimum required rate of return.

With the corrected formulation of maximizing R subject to (2) and (4), the Lagrangian becomes

$$\mathcal{L} = R(Q) - \lambda[R - wL - sK] + \gamma[R - wL - \rho K], \quad (5)$$

and the first-order conditions

$$R'Q_L(1 - \lambda + \gamma) + (\lambda - \gamma)w = 0 \quad (6)$$

$$R'Q_K(1 - \lambda + \gamma) + \lambda s - \gamma\rho = 0, \quad (7)$$

with the complementary conditions

$$\lambda[R - wL - sK] = 0 \quad \text{and} \quad \gamma[R - wL - \rho K] = 0, \quad (8)$$

and from the form of the constraints we know that $\lambda, \gamma \geq 0$. Moreover, where $\gamma = 0$, the problem reduces to that considered by Bailey and Malone and by the same argument as theirs it may be deduced that, in this case $\lambda < 1$. In other words where $\gamma = 0$ we have $0 \leq \lambda < 1$.

⁵ Bailey and Malone also consider other forms of regulation but these are not discussed here.

⁶ In [1].

⁷ See [3].

⁸ The constraint is formulated in terms of a minimum return on capital, which seems more plausible than the absolute profit constraint of Baumol's original article. The effect of this change in specification is discussed below.

■ In their discussion of the problem, Baumol and Klevorick⁹ generously interpret Bailey and Malone as having implicitly assumed that $s = \rho$ (the rate of return set by regulators equals the minimum profit acceptable to the firm). In that case we have, in effect, a single constraint which has to hold with equality, and the input combination is given as they say by

$$\frac{Q_L}{Q_K} = \frac{w}{s}.$$

There is, however, no reason to suppose that s will equal ρ . It is quite reasonable to suppose, for example, that the minimum rate of return the sales-maximizing firm has to achieve is equal to the cost of capital ($\rho = i$). In that case the assumption that the fair rate (s) exceeds i is sufficient to rule out the implicit assumption of Bailey and Malone.¹⁰

In the case where $\rho = i$ the possible solutions are set out in Table 1. In Case II we have the Bailey-Malone solution with under-capitaliz-

TABLE 1

	$\lambda > 0$	$\lambda = 0$
$\gamma > 0$	NOT POSSIBLE SINCE $s > \rho$	CASE I $R = wL + \rho K$ $R' > 0$ SINCE $\gamma > 0$ $Q_L / Q_K = w / \rho = w / i$
$\gamma = 0$	CASE II $R = wL + sK$ $R' < 0$ SINCE $\frac{\lambda}{1 - \lambda} > 0$ $Q_L / Q_K = w / s < w / i$	CASE III NEITHER CONSTRAINT BINDS $R' = 0$ SINCE $Q_K > 0$, $Q_L > 0$ BY ASSUMPTION INPUT COMBINATION INDETERMINATE

ation (in the sense defined earlier) but this is not the only possible solution. Moreover, there are good reasons to suppose that it is less likely to occur than Case I, where there is no inefficiency in the input combination chosen. It is clear, for example, that if the demand for the firm's product were everywhere price elastic, then Cases II and III would be ruled out (since $R' > 0$). An alternative approach is that adopted originally by Baumol (extended by Haveman and DeBartolo¹¹ and Kafoglis and Bushnell¹²) of introducing advertising. If we denote advertising by A , the Lagrangian becomes

$$\mathcal{L} = R - \lambda[R - wL - A - sK] + \gamma[R - wL - A - \rho K], \quad (9)$$

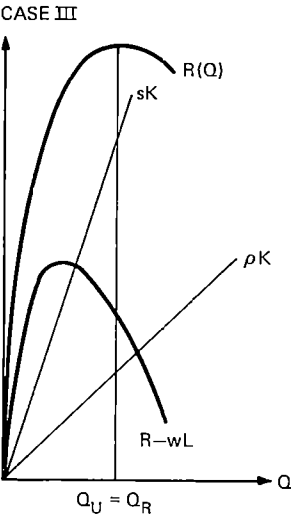
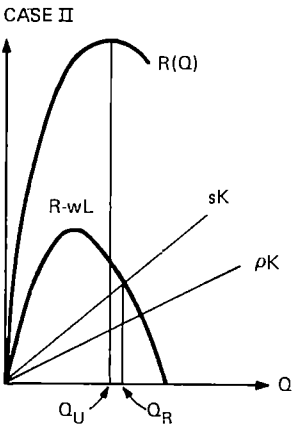
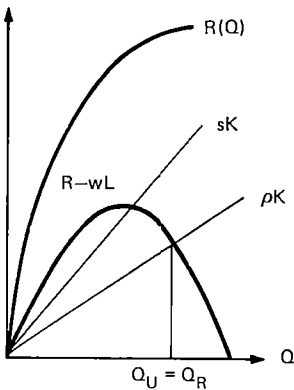
where $R = p(Q, A)Q$. The first-order condition with respect to A is

$$\frac{\partial R}{\partial A} (1 - \lambda + \gamma) + \lambda - \gamma \leq 0,$$

⁹ In [5].
¹⁰ If $s < \rho$, the firm would have to cease operation and this possibility is not considered.
¹¹ In [7].
¹² In [9].

3. Interpretation of the results

FIGURE 1
CASE I



with equality if $A > 0$. If we assume $\partial R/\partial A > 0$, then this rules out a solution with $\gamma = 0$ (given that $\frac{\lambda}{1-\lambda}$ is positive) and the only possible solution is Case I. It follows, therefore, that we are able to rule out the Bailey-Malone result on the assumption that either $\partial R/\partial Q > 0$ or $\partial R/\partial A > 0$, the latter appearing to be a relatively weak requirement. Moreover, we may note that under these conditions the firm's behavior is unaffected by regulation—the regulatory constraint is not binding. These results may be summarized as follows:

Proposition 1: If $\rho = i$, and either $\partial R/\partial Q > 0$ or $\partial R/\partial A > 0$, then the sales-maximizing firm's behavior is unaffected by regulation and there is no inefficiency in the input combination chosen.

In the case where $\rho \neq i$ (and $\rho < s$) it is clear that Case I may lead to the choice of inefficient input combinations, and the conditions on $\partial R/\partial Q$ and $\partial R/\partial A$ are no longer sufficient to rule out the Bailey-Malone result. If $\rho > i$, then Case I involves undercapitalization. However, this inefficiency has nothing to do with regulation as suggested by Bailey and Malone and the inefficiency arises equally in the unregulated firm. An unregulated sales-maximizing firm chooses a lower (higher) capital-labor ratio for a given output than a profit-maximizing firm accordingly as ρ is greater (less) than i . This result appears to conflict with the statement by Baumol¹³ that the sales-maximizing firm chooses the same input combination as a profit-maximizing firm but is explained by the form of the minimum profit constraint, which in Baumol's case is written as $R - wL \geq P^*$. It does not appear to have been noted in the literature that Baumol's statement depends crucially on the form of the constraint.¹⁴

Proposition 2. If $i \neq \rho < s$, and either $\partial R/\partial Q > 0$ or $\partial R/\partial A > 0$, then the sales-maximizing firm exhibits undercapitalization (where $\rho > i$) or overcapitalization where ($\rho < i$), but this is not attributable to regulation, which has no effect on the firm's behavior.

4. A geometric treatment

■ The different solutions are illustrated in Figure 1¹⁵ (where for ease of geometric representation it is assumed that there are fixed input coefficients). Denoting by Q_U the output in the unregulated case and by Q_R the output when regulated, we can see that the only case in which regulation makes any difference is Case II. We have seen that this case can be ruled out by $\partial R/\partial Q > 0$ (or by $\partial R/\partial A > 0$ which cannot be represented on this diagram). Moreover, it is clear that if we considered an output-maximizing firm (one of the other objectives considered by Bailey and Malone), then the only possible solution in the unregulated case is where the constraint is binding. Finally we may note that where $\partial R/\partial Q > 0$, the formulation by Bailey and Malone of maximizing R subject to the *maximum* profit constraint

¹³ In [3].

¹⁴ In [4], Notes to Chapter 7, Baumol replaces the fixed profit constraint by a constraint related to total costs and argues that "none of the previous analysis and its conclusions is affected." What he does not consider is a constraint in which the inputs are treated asymmetrically—as in the rate of return on capital constraint discussed here.

¹⁵ On p. 285.

(2) is clearly absurd—without the minimum profit constraint (4), no optimal policy for the firm exists.

■ In their survey article, Baumol and Klevorick refer to the examination of the sales-maximization case by Bailey and Malone as “one of a number of attempts to show that the A-J [Averch-Johnson] conclusions lack general validity.”¹⁶ The purpose of this note has been to show that this conclusion is correct for rather different reasons than they suggested. Under relatively weak assumptions, regulation has no effect on the behavior of the sales-maximizing firm—a result which, while quite intuitive, is in striking contrast to the Averch-Johnson findings for the case of profit-maximization. For this reason, regulation does not have the implications for resource allocation indicated by Bailey and Malone and under the conditions described above, any tendency to undercapitalization has nothing to do with the fair-rate-of-return constraint.

¹⁶ See [5].

References

1. AVERCH, H. AND JOHNSON, L. L. “Behavior of the Firm under Regulatory Constraint.” *The American Economic Review*, Vol. 52, No. 5 (December 1962), pp. 1053–1069.
2. BAILEY, E. E. AND MALONE, J. C. “Resource Allocation and the Regulated Firm.” *The Bell Journal of Economics and Management Science*, Vol. 1, No. 1 (Spring 1970), pp. 129–142.
3. BAUMOL, W. J. “On the Theory of Oligopoly.” *Economica*, Vol. 25, No. 3 (August 1958), pp. 187–198.
4. ———. *Business Behavior, Value and Growth*. New York: Harcourt Brace, 1967.
5. ——— AND KLEVORICK, A. K. “Input Choices and Rate of Return Regulation: an Overview of the Discussion.” *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 162–190.
6. EDELSON, N. M. “Resource Allocation and the Regulated Firm: A Reply to Bailey and Malone.” *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 374–378.
7. HAVEMAN, R. AND DEBARTOLO, G. “The Revenue Maximization Oligopoly Model: Comment.” *The American Economic Review*, Vol. 58, No. 5 (December 1968), pp. 1355–1358.
8. KAFOGLIS, M. “Output of the Restrained Firm.” *The American Economic Review*, Vol. 59, No. 4 (September 1969), pp. 583–589.
9. ——— AND BUSHNELL, R. C. “The Revenue Maximization Oligopoly Model: Comment.” *The American Economic Review*, Vol. 60, No. 3 (June 1970), pp. 427–428.
10. ZAJAC, E. E. “A Geometric Treatment of Averch-Johnson’s Behavior of the Firm Model.” *The American Economic Review*, Vol. 60, No. 1 (March 1970), pp. 117–125.

5. Conclusions

Resource allocation and the regulated firm: comment on the comments

Elizabeth E. Bailey

Economics Modeling Group
Bell Telephone Laboratories

The revenue-maximizing firm presumably has an internal return requirement that is at least as large as the maximum return permitted by a regulatory agency. In this case, which is the only one that seems economically meaningful, the undercapitalization result of Bailey-Malone holds throughout the elastic region of the revenue curve.

1. Internal constraint

■ The behavior of a regulated sales-maximizing firm has a number of interesting features that were not fully recognized by Bailey and Malone.¹ Baumol and Klevorick pointed out that our stipulation that equality holds in the regulatory constraint depends on an implicit assumption that the managers of the firm wish to earn at least the rate of return on capital permitted by the regulatory agency. In their words:

"The discussion implicitly assumes that the profit constraint of the regulatory model also serves, with the inequality reversed, as the *minimum* profit requirement for the sales-maximization hypothesis."²

This criticism is both pertinent and correct. Notice, however, that expressed in terms of profit, the regulatory constraint is $\pi \leq (s - i)K$ (for s = fair return, i = cost of capital, $s > i$, and K = capital), whereas the profit constraint ordinarily used in a sales-maximization model is $\pi \geq \pi^*$ (for π^* , a fixed dollar profit).

Edelson, like Baumol and Klevorick, recognized that undercapitalization by the regulated sales-maximizing firm follows directly from the assumption that the regulatory constraint is binding. However, he made a further assertion:

Suppose we begin with the conventional model of a sales- or output-maximizing firm, *viz.*, that management maximizes its objective function subject to some self-imposed lower bound on the profit rate. If this acceptable rate of return on investment is no greater than the "fair" rate, regulation will have no effect on the firm's output or costs.³

Thus, Edelson is arguing that it is possible to suppose that the managers of the firm may have a target rate of return which is less than the regulatory rate. He notes that if the firm's internal constraint is at a level i^* , $i < i^* < s$, then the internal constraint, *not the regu-*

The author is grateful to William J. Baumol, David Dayan, and Gerald R. Faulhaber for their very helpful suggestions on the first draft of these comments. Views expressed in the comments are those of the author. They should not be interpreted as reflecting the official opinion or policy of Bell Telephone Laboratories or of the American Telephone and Telegraph Company.

¹ In [4].

² In [6], p. 181.

³ In [8], p. 376.

latory constraint, is what serves to distort resource allocation. When $i^* = i < s$, Edelson asserts that there will be minimum-cost production. When $i^* \geq s$, the Bailey-Malone results apply.

Atkinson and Waverman reiterate the Edelson conclusions in their Propositions 1 and 2, however, with the qualification that the statements only hold if marginal revenue is positive at the optimal solution to the constrained system. Unfortunately, Atkinson and Waverman choose to ignore the third result. This exclusion leads to incorrect statements such as:

We are able to rule out the Bailey-Malone result on the assumption that . . . $\delta R/\delta Q > 0$ [marginal revenue is positive].⁴

Simply by recognizing that the firm can wish to earn at least as much return as the regulator permits, they would have had to record in the upper left-hand corner of their Table 1⁵ the theoretical conclusion that regulation can be effective and can entail undercapitalization, as Bailey and Malone asserted. Thus, Atkinson and Waverman do not raise a legitimate criticism on the undercapitalization result.

The legitimate issue that is raised by Edelson and by Atkinson and Waverman concerns the relative importance of the two alternative stipulations about the firm's internal profit constraint. The other authors emphasize the case where the firm has a lower target return than the regulatory fair rate, while Bailey-Malone restricted the analysis to the case where the internal target is at least as large as the fair return. I will argue now that the Bailey-Malone approach is the relevant one, and that the other authors' emphasis is misplaced.

The reasons are most easily highlighted by using a simple numerical example. All the discussants are adhering to the Averch and Johnson assumptions that the fair return exceeds the market cost of capital and that the regulated firm earns a lower return than would a profit-maximizing monopolist.⁶ Let us say that the profit maximizer's return is 15 percent, the target return under a sales objective is 12 percent, the fair return is 10 percent, and the cost of capital is 9 percent. In this case regulation lowers the actual return of the firm from the 12- to the 10-percent level and effectively binds the firm below its target return. Furthermore, the firm is depicted as earning the regulatory rate rather than a lower rate. This is a case in which actual earnings are at the maximum level permitted.

Under the Edelson or Atkinson and Waverman scheme, the unregulated firm has a target return of 9.5 percent, a return which is *less than* the 10-percent rate that the regulator deems fair. Under this scheme the firm's managers are electing a lower return than that stipulated by the regulator, a situation that would almost certainly not be looked upon favorably by its stockholders. Furthermore, since binding regulation would mean a return of 10 percent rather than 9.5 percent, society would have achieved a lower price in the absence of regulation than with effective regulation.

⁴ In [1], p. 286.

⁵ *Ibid.*, p. 285.

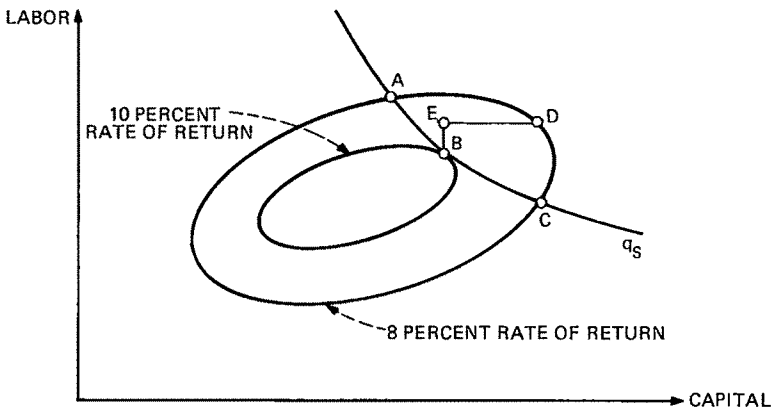
⁶ See [2]. It should be mentioned, however, that many observers of the regulatory scene have argued that the fair return is set equal to, not larger than, the market cost of capital, and indeed that in times of inflation fair return may even drop below market cost of capital, at least in the short run. For a recent review of some of these issues see Johnson [9].

2. Revenue maximized

■ If this were the entire story, the undercapitalization result derived in Bailey and Malone would hold for the only case that we claim is relevant. However, the story does not end here. Both Bailey-Malone and Edelson failed to take account of the possibility that the revenue objective may actually be achieved at the optimal solution to the constrained system. This case is probably of minor importance in the real world, where the existence of advertising and quality-improving alternatives ordinarily ensures a positive marginal revenue. However, in the standard model of the regulated firm in which there is a single output produced with labor and capital inputs, it is possible, as Atkinson and Waverman say, that demand and technology conditions are such that the revenue-maximizing output yields a return on capital larger than the regulatory rate.

Figure 1 illustrates the situation. The revenue (sales)-maximizing isoquant is labelled q_s . Two of the firm's isorate-of-return contours

FIGURE 1



are also depicted. As the tangency at point B illustrates, the sales-maximizing output is the maximum output that can be produced at a rate of return $= r > s$. When the regulatory constraint binds the firm to the rate of return $= s$ contour, it is clear that the firm can still satisfy its revenue maximization objective if it shifts its combination of labor and capital and operates at either point A or point C . Hence, Atkinson and Waverman are correct in the indeterminacy in resource allocation, which they record as Case III in their Table 1.⁷

However, Atkinson and Waverman make an error in Case II of their Table 1.⁸ The firm will *never* find it optimal to operate in the inelastic region of demand, for such a move lowers revenue, and revenue is what the firm is maximizing. In terms of Figure 1, point D lies on an isoquant $q > q_s$ and hence is distinctly an inferior choice to either of points A or C , since revenue is higher at these latter points than it is at the former.

There is another alternative that is presumably open to the firm in the situation depicted in Figure 1. It could operate off its production possibilities boundary.⁹ This would occur, for example, if the

⁷ In [1], p. 285.

⁸ *Ibid.*

⁹ A general discussion of operation off the production frontier in regulatory models is given in Bailey [3]. See also Dayan [7].

firm used the labor and capital indicated at point *D* but continued to produce the revenue-maximizing output. There would then be a wasteful use of *ED* worth of capital and *EB* worth of labor. Costs are not minimized because the firm could produce more, given the labor and capital resources expended. In terms of familiar theoretical models, the firm could adopt managerial emoluments of the Williamson variety¹⁰ of an amount *EB*, and it could goldplate by an amount *ED*.¹¹ The particular solution that the firm adopts can only be determined either if more is known about the secondary objectives of the firm, or if something is known about the regulator's ability to detect and disallow the various types of nonminimum cost expenditure. Thus, the indeterminacy noted by Atkinson and Waverman can be interpreted somewhat more broadly than the overcapitalization and/or undercapitalization alternatives usually cited in regulatory models and, contrary to what Edelson asserts,¹² can involve waste as a part of a solution to the system.

■ Perhaps the most valuable purpose that has been served by all this discussion is that the distinction between profit constraints and rate-of-return constraints has been clarified more fully. In the case of the profit constraint, the firm chooses the point of largest output on a given isoprofit contour. As long as marginal revenue is positive at the solution to this model, the usual result of Baumol¹³ holds: the firm operates on the efficient path.

In the case of the rate-of-return constraint, the firm chooses the point of largest output on a given isorate-of-return contour. The crucial feature is that the point at which output is maximized on such a contour is not a point of minimum-cost production. The explanation is essentially this. The tangent line between an isoquant and an isoprofit contour has a slope equal to the ratio of the factor prices ($-dL/dK = i/w$, where i = cost of capital and w = wage rate). Hence, the point of highest output on a given isoprofit contour is always one at which costs are minimized. However, the tangent line between an isoquant and an isorate-of-return contour has a slope equal to r/w , where r = given rate of return. Thus, if the return r is larger than the cost of capital, the point of highest output on a given isorate-of-return contour is not associated with minimum-cost technology, but rather is one at which the firm undercapitalizes.¹⁴

These observations prove crucial for the classical sales-maximizing model as well as for the model of the regulated firm with a sales objective. If the classical constraint had specified a given rate of return rather than a given level of profit, then as long as the constraint binds the firm below the revenue-maximizing output, the firm comes closest to its sales objective by minimizing the effect of the constraint, that is, by using too little capital. Thus a rate-of-return constraint causes a misallocation of resources in the classical sales-maximizing model as well as in the regulatory model presented in Bailey-Malone.

3. Profit constraints and rate-of-return constraints

¹⁰ In [10].

¹¹ See Zajac [11].

¹² In [8], p. 375.

¹³ In [5].

¹⁴ For a more detailed derivation, see Bailey [3].

References

1. ATKINSON, A. B. AND WAVERMAN, L. "Resource Allocation and the Regulated Firm: Comment." *The Bell Journal of Economics and Management Science*, Vol. 4, No. 1 (Spring 1973), pp. 283-287.
2. AVERCH, H. AND JOHNSON, L. L. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol. 52, No. 5 (December 1962), pp. 1053-1069.
3. BAILEY, E. E. *Economic Theory of Regulatory Constraint*. Heath-Lexington, forthcoming.
4. ——— AND MALONE, J. C. "Resource Allocation and the Regulated Firm." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 1 (Spring 1970), pp. 129-142.
5. BAUMOL, W. J. *Business Behavior, Value and Growth*. New York: Harcourt-Brace, 1967.
6. ——— AND KLEVORICK, A. K. "Input Choices and Rate-of-Return Regulation: An Overview of the Discussion." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 162-190.
7. DAYAN, D. "Vertical Integration and Monopoly Regulation." Unpublished Ph.D. dissertation, Princeton University, December 1972.
8. EDELSON, N. M. "Resource Allocation and the Regulated Firm: A Reply to Bailey and Malone." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 374-378.
9. JOHNSON, L. L. "Behavior of the Firm under Regulatory Constraint: A Re-assessment." *The American Economic Review*, Vol. 63, No. 2 (May 1973).
10. WILLIAMSON, O. E. *The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm*. Englewood Cliffs, N. J.: Prentice-Hall, 1964.
11. ZAJAC, E. E. "Note on 'Gold Plating' or 'Rate Base Padding'." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 1 (Spring 1972), pp. 311-315.

The profit-performance effects of the separation of ownership from control in large U. S. industrial corporations

John Palmer

Assistant Professor of Economics
The University of Western Ontario

After a survey of the empirical studies of Monsen, Chiu, and Cooley, of Kamerschen, and of Larner on the effects of the separation of ownership from control, this paper presents a study which shows that among firms with a high degree of monopoly power, management-controlled firms report significantly lower profit rates than owner-controlled firms. The works of Bain, Mann, and Shepherd and a 1967 FTC study supply information on barriers to entry, which more aptly capture the effect of separation of ownership and control. The findings confirm Hall and Weiss' conclusion that there is less variation in profit rates among large firms than among small ones.

■ Empirical evidence on the effects of the separation of ownership from control has, to date, been mixed. The results of Monsen, Chiu, and Cooley¹ (M-C-C) generally support the hypothesis that management-controlled firms are considerably less profitable than owner-controlled firms. Kamerschen,² however, found no difference between the average reported profit rates of management-controlled firms and owner-controlled firms. Larner³ discovered only slightly lower profit rates reported by management-controlled firms than by owner-controlled firms. This paper, using a broader sample, different independent variables, and different statistical techniques, presents evidence that among firms with a high degree of monopoly power, management-controlled firms *do* report significantly lower profit rates than owner-controlled firms.

■ M-C-C's work can be criticized on two counts. First, while M-C-C did include an industry variable in their statistical tests to allow for different effects in different industries, the variable used had twelve classes, permitting no distinction between different types of industries with different structures. Second, by not correcting for heteroscedasticity in their tests, M-C-C used biased estimates of class variances which distort their results.

Kamerschen's results can be criticized on many grounds. First, his inclusion of the variable, "change in control from 1929 to 1963," is

Helpful comments from Professors L. B. Fletcher and R. W. Mensing and from the Editor of The Bell Journal are gratefully acknowledged, although final responsibility for any errors which may remain rests with the author.

¹ See [10].

² See [4].

³ See [5].

1. Introduction

2. Review of previous studies

not theoretically justifiable because of the long period of time which elapsed between the two dates during which other events (e.g., depression and war) took place, and because a test of the effect of a change in control should involve a time series analysis of those firms which experienced such a change. Second, including industry growth rates from 1947–1953 to 1957–1960 as an independent variable along with the growth rates from 1957–1960 to 1960–1964 in a study of average profit rates from 1959–1964 is similarly questionable on theoretical grounds and on account of the potential for multicollinearity introduced into the regression. Third, there is also a strong possibility of multicollinearity between the concentration ratios and industry barriers to entry, both of which were introduced as independent variables in the regressions. Fourth, Kamerschen, too, did not test for or correct for heteroscedasticity in his tests.

Of the three studies referred to in the introduction, that of Larner is perhaps the most appealing. He did introduce a measure of monopoly power (four-firm concentration ratios), and he did attempt a correction, though admittedly imperfect,⁴ for heteroscedasticity. Perhaps the major reason for the differences between his results and those of this study is the allowance here for an interaction effect between monopoly power and the type of control of the firm. Simply including both variables in the regressions, as Larner did, assumes that the effect of the separation of ownership from control should be the same for all degrees of monopoly power and forces such an assumption onto the statistical results. If, however, the managers of management-controlled firms with a high degree of monopoly power, because of their relative freedom from the rigors of competition, have more discretion in the use of their firms' potential profits than their counterparts in firms with a low degree of monopoly power, such an assumption is incorrect, and an interaction term should be included.

3. The Palmer study

■ The dependent variable used was the average rate of return on net worth from 1961 to 1965 and from 1966 to 1969 for each of the 500 largest corporations in 1965 listed in *Fortune*. The three independent variables used to "explain" the dependent variable were type of control of the firm, monopoly power of the firm, and size of the firm.

Each firm was assigned to one of the following three classes according to its type of ultimate control in 1965 and in 1969:

- (1) Strong owner control—one party owned 30 percent or more of the outstanding common stock of the corporation.
- (2) Weak owner control—one party owned between 10 percent and 29.9 percent of the outstanding common stock of the corporation.
- (3) Management control—no single party owned 10 percent or more of the outstanding common stock of the corporation.

As in studies by Larner and M-C-C, "party" is interpreted to cover families, the board of directors, and/or some close-knit group of business associates. The owner-control category was subdivided into weak owner control and strong owner control to allow for the

⁴ *Ibid.*, pp. 28–29.

possibility discussed by M-C-C that the intermediate type of control (weak owner control) might contain some firms which are more nearly management-controlled and some other firms which are more nearly owner-controlled. If this intermediate group is mixed, then a comparison of just the two extreme groups should provide a clearer test of the effect of the separation of ownership from control. Admittedly, the 30-percent cut-off point between weak and strong owner control is somewhat arbitrary, but it was set at this rather high level to obtain a more definitive distinction between management and strong owner control.⁵

Barriers to entry (*BTE*) were used to estimate the monopoly power of each firm for three reasons: (1) they were more appealing theoretically than concentration ratios in that they take account of potential competition; (2) in Kamerschen's study *BTE* were statistically significant and concentration ratios were not; and (3) in a study by Hall and Weiss⁶ and in the previously mentioned work by Lerner, concentration ratios were not statistically significant. It was felt that if monopoly power does significantly affect a firm's profit rate, a measure should be used which would capture this effect. *BTE* have been shown by Mann⁷ to capture the effect, and hence were decided upon for this study.

Whenever possible, previous studies were relied upon to provide information on *BTE*. The four primary sources were Bain, Mann, and Shepherd.⁸ Supporting evidence was frequently drawn from a Federal Trade Commission (FTC) study done in 1967.⁹

Bain and Mann discuss four basic types of barriers to entry. These are:

- (1) *Scale barriers*—The cost curves for firms in a particular industry may be downward sloping until a very large rate of output is reached. This rate of output may be so large relative to the total demand for the product that only a few firms can profitably exist in the industry. Entry by another firm would reduce the market shares of all firms and raise their average costs so that none of them could be profitable. Potential entrants, recognizing this state of affairs, will try to avoid unprofitable entry, but the incumbents may still be enjoying monopoly profits.
- (2) *Product differentiation barriers*—If consumers are strongly attached to existing brands of a product, either because they find the relevant information difficult to acquire or comprehend or because of large advertising expenditures, entrants to this industry may be forced to sustain burdensome advertising expenditures if they wish to capture a profitable share of the market.
- (3) *Absolute cost barriers*—If one firm or a few firms control necessary patents or natural resources or have special technical secrets, entry into this industry will be effectively impeded.

⁵ The sources and data for the independent variables are presented in an unpublished appendix which is available from the author on request.

⁶ See [2].

⁷ In [7, 8].

⁸ See [1], [7, 8], and [13], respectively.

⁹ See [16].

- (4) *Capital barriers*—If undertaking a business venture, a firm will require capital in the form of plant, equipment, and inventories. If the capital market is imperfect, large capital requirements in an industry may be a barrier to entry, since small firms may be unable to obtain investment funds because of rationing or may be forced to pay significantly higher rates of interest for the investment funds they do get.

The monopoly power of each firm was measured by first estimating the height of the *BTE* (very high, substantial, or moderate-to-low) into each industry in which each firm operated. Then those industries with very high barriers were assigned the value of 1.0, substantial 0.5, and moderate-to-low 0.0. To allow for diversification across industry lines, the percentage of each firm's sales in each industry was multiplied by the weight of that industry's *BTE*, and the sum was used to indicate each firm's degree of monopoly power. On the basis of this weighted sum, each firm was assigned to one of the following three categories:

- (1) High monopoly power: the weighted sum ≥ 0.667 ;
- (2) Medium monopoly power: $0.667 > \text{the weighted sum} > 0.333$;
- (3) Low monopoly power: $0.333 \geq \text{the weighted sum}$.

It should be noted that not all economists agree on the appropriateness of *BTE* as a measure of monopoly power, particularly product differentiation barriers¹⁰ and capital barriers.¹¹ It may be, too, that *BTE* and concentration ratios do not have significantly different effects on profit rates.¹² The point to be emphasized here, however, is not that *BTE* are necessarily the best measure of monopoly power, but that *BTE* interacting with the type of control give better results than previous studies of the separation of ownership from control.

Firm size was determined from total sales figures reported in *Fortune* in 1965 and 1969. Each firm was assigned to one of four classes—the 125 largest firms, the next 125 firms, etc., for both 1965 and 1969. Class measures were used for the type of control and monopoly power variables to help mask out some of the problems due to errors in measurement.¹³ The same procedure was carried over to the size variable to facilitate the statistical comparison technique employed.

4. Statistical tests

■ The existence of heteroscedasticity can, if severe, provide biased estimates of the variances of regression coefficients and, as a result, render tests of significance meaningless. Consequently, statistical techniques must be utilized which correct for the problem. Hall and Weiss, in their cross-sectional analysis of firm size and profitability found heteroscedasticity present and corrected for it to some extent. Larner applied the same technique (weighted regressions) in his study of the separation of ownership from control.

¹⁰ See Telser [15].

¹¹ See Stigler [14].

¹² See Rhoades [12].

¹³ See Johnston [3], pp. 164-165.

To determine whether heteroscedasticity might be present in this study, a factorial analysis was performed, based on the range of the dependent variable within each of the 36 cells (three types of control, three classes of monopoly power, and four size classes). Heteroscedasticity was indicated for both the type of control and size (see Table 1). Its presence not only confirms the finding of Hall and Weiss

TABLE 1
AOV TABLE FOR TESTS OF HETEROSCEDASTICITY AMONG AVERAGE
REPORTED PROFIT RATES FOR DIFFERENT TREATMENTS

SOURCE	d.f	SS	F
MEAN	1	15273.25	
TREATMENTS	23		
1. TYPE OF CONTROL	2	1939.55	23.07 ^a
2. SIZE	3	714.53	5.67 ^a
3. MONOPOLY POWER	2	274.70	3.27
1 x 2	6	295.21	1.17
1 x 3	4	90.77	0.54
2 x 3	6	804.73	3.19
ERROR	12	504.50	
AVERAGE RANGES:			
MANAGEMENT CONTROL		30.43	
WEAK OWNER CONTROL		18.57	
STRONG OWNER CONTROL		12.80	
LARGEST SIZE CLASS		13.21	
SECOND SIZE CLASS		23.35	
THIRD SIZE CLASS		24.73	
FOURTH SIZE CLASS		21.10	

^aSIGNIFICANT AT THE FIVE - PERCENT LEVEL

that there is less variation in profit rates among large firms than among small firms,¹⁴ but also has some interesting independent implications concerning the effect of the separation of ownership from control. One possible explanation of this phenomenon is that managers of management-controlled corporations have different utility functions and different estimates of the weaknesses of stockholder control, leading to different amounts of effort by them to maximize reported profits. In owner-controlled firms, however, the managers have less scope for discretionary action, even if their utility functions are different, leading to reported profits by them which are consistently closer to the maximum level.¹⁵

The problem of heteroscedasticity was compensated for by the use of statistical comparisons (or contrasts) between different classes or combinations of classes. Each comparison could then be tested by subdividing the total error sum of squares into just that portion due to each comparison.¹⁶ Because this comparison technique provides

¹⁴ In [2].
¹⁵ Put differently, heteroscedasticity with respect to the type of control classes would imply that the average reported profit rates for each of the management-controlled firms would be more widely dispersed about the mean of that category than would the average reported profit rates for each of the owner-controlled firms about the mean of their group.
¹⁶ For a detailed discussion of the technique, see Ostle [11], p. 337. It should be noted that the previous test for heteroscedasticity is not strictly legitimate because it is based on cell ranges; all that can be said is that the test indicates a strong likelihood of the presence of heteroscedasticity. The comparison tests used in this study will take account of heteroscedasticity if it is present, but will not be affected if it is not.

unbiased tests of significance, it involves a better correction for heteroscedasticity than that used by Larner. Furthermore, it has the advantage that it allows direct tests of all possible interactions between the type of control and monopoly power.

5. Results and conclusions

■ The average reported profit rates for firms in each possible category of control and monopoly power are shown in Table 2, and

TABLE 2

AVERAGE REPORTED PROFIT RATES, BY TYPE OF CONTROL AND DEGREE OF MONOPOLY POWER

MONOPOLY POWER	TYPE OF CONTROL		
	MANAGEMENT	WEAK OWNER	STRONG OWNER
HIGH	11.41	14.53	14.77
MEDIUM	10.26	12.12	9.95
LOW	9.89	9.70	10.59

the results of some of the comparisons involving these averages are presented in Table 3.¹⁷ The most interesting result is that only among firms with a high degree of monopoly power is there a significant difference between the average profit rates of management-controlled firms vs. owner-controlled firms; all of the other comparisons involving both the type of control and the degree of monopoly power are statistically insignificant at the 5-percent level. Among those firms with a high degree of monopoly power, the average profit rate of management-controlled firms was significantly lower than that for both classes of owner-controlled when the latter were grouped together, though the difference between management- and weak owner-controlled firms alone was not significant. These results are not sur-

TABLE 3

COMPARISONS OF AVERAGE REPORTED PROFIT RATES FOR DIFFERENT TYPES OF CONTROL, HOLDING MONOPOLY POWER CONSTANT

COMPARISON	d.f.	F
LOW MONOPOLY POWER		
MANAGEMENT VS. WEAK OWNER CONTROL	1	0.059
MANAGEMENT VS. STRONG OWNER CONTROL	1	0.799
WEAK OWNER VS. STRONG OWNER CONTROL	1	0.902
MEDIUM MONOPOLY POWER		
MANAGEMENT VS. WEAK OWNER CONTROL	1	7.508
WEAK OWNER VS. STRONG OWNER CONTROL	5	6.272
HIGH MONOPOLY POWER		
MANAGEMENT VS. ALL OWNER CONTROL	5	13.039 ^a
MANAGEMENT VS. WEAK OWNER CONTROL	3	5.109
MANAGEMENT VS. STRONG OWNER CONTROL	7	15.464 ^a
^a SIGNIFICANT AT THE FIVE-PERCENT LEVEL (d.f. ESTIMATED USING SATTERTHWAITHE'S APPROXIMATION)		

¹⁷ None of the comparisons of the effect of the type of control, holding size classes constant, was significant.

prising if a manager is viewed as operating under two constraints—one from stockholders and one from competition in the market. Only if *both* constraints are very weak would one expect a significant divergence from profit-maximizing behavior, as was observed.

Next, just those management-controlled firms with a high degree of monopoly power, but reporting relatively low profit rates, were examined to see whether there might exist other reasons for their poorer performances. In some cases other reasons were immediately apparent. For example, Western Electric may well be subject to at least second-order regulation and hence be required to earn lower profit rates.¹⁸ Also Allis Chalmers and J. I. Case, producing farm implements, are in a relatively declining industry and would be expected, therefore, to report lower profit rates. For the most part, however, no clear pattern emerged to explain the lower profit rates. The list of firms in this category included Chrysler, Westinghouse, Boeing, General Dynamics, Sperry Rand, Anaconda, Colgate Palmolive, American Brands, Honeywell, National Distillers and Chemical, Babcock and Wilcox, Philip Morris (from 1961–1965), Northrup, Parke Davis, ITE Imperial, and Curtiss Wright. But each of these firms had high-monopoly-power counterparts in similar industries with better performances, such as General Motors, RCA, Lockheed (from 1961–1965), Lorillard (from 1961–1965), IBM, Pfizer, Bristol Meyers, Emerson Electric, Square D, Rohr, Cessna, U. S. Smelting (from 1966–1969), Seagram, and Englehard (from 1966–1969). For the most part, it appears that no distinguishing feature is present to explain the different profit rates reported by differently controlled firms. Certainly, to some extent the results reported in Tables 2 and 3 must be qualified because of the inclusion of firms like Western Electric, Allis Chalmers, and J. I. Case in the sample, but the large number of firms for which exceptions cannot be easily made leads one to believe that the results are essentially correct.

As a final point, mention must be made of recent studies indicating that executive compensation packages in stock options, etc., provide some incentive for managers to increase the profitability or present value of their corporations.¹⁹ While these studies may be correct, one must be careful not to infer too much from them, especially for corporations with a high degree of monopoly power, for the evidence presented here suggests that such compensatory schemes do not create an absolute constraint on managerial behavior but are instead just another argument in managers' utility functions which might be sacrificed to some extent for attainment of other objectives.

Appendix

■ Whenever possible, previous studies were relied upon to provide information on barriers to entry to these industries. The primary sources were Bain, Mann, and Shepherd. Below are listed industries with very high (VH), substantial (S), or moderate-to-low (M-L) barriers to entry, respectively. The codes in parentheses following each industry refer to the source(s) in which the classification was made. These sources and their codes are:

Barriers to entry to various industry

¹⁸ Or the firm may successfully have hidden higher potential profits from the regulators.

¹⁹ In Lerner [5], chapter 4; Lewellen [6]; and Masson [9].

B—Bain—[1]
 MR—Mann—[8]
 MW—Mann—[7]
 Sh—Shepherd—[13]
 P—Palmer—this study

An asterisk following an industry indicates a disagreement among sources as to the appropriate class for that industry. Resolutions of these disagreements along with explanations of all classifications coded "P" are available from the author.

□ **Industries with very high barriers to entry :**

- | | |
|--|---|
| (1) automobiles (B, MR, MW, Sh) | (21) computing and related machines (Sh) |
| (2) farm machinery and equipment, integrated line or heavy equipment (*) | (22) transformers (Sh) |
| (3) cereal (*) | (23) switchgear and switchboards (Sh) |
| (4) chewing gum (*) | (24) electric motors and generators (Sh) |
| (5) distilled liquor (B, MR, MW, Sh). | (25) industrial controls (Sh) |
| (6) cigarettes (*) | (26) carbon and graphite products (MW, Sh) |
| (7) pulp mills (Sh) | (27) buses (Sh) |
| (8) newspapers (Sh) | (28) electric lamps (Sh) |
| (9) sulphur (MR, MW) | (29) telephone and telegraph apparatus (Sh) |
| (10) nickel (MR, MW) | (30) aircraft engines and parts (Sh) |
| (11) copying equipment (Sh) | (31) aircraft (Sh) |
| (12) soft drink flavorings (MW) | (32) shipbuilding and repairing (Sh) |
| (13) ethical drugs (MR, Sh) | (33) locomotives and parts (Sh) |
| (14) proprietary drugs (*) | (34) photographic equipment (Sh) |
| (15) soap and other detergents (*) | (35) aerospace and defense (P) |
| (16) explosives (Sh) | (36) iron ore (P) |
| (17) flat glass (MR, MW, Sh) | (37) copper ore (P) |
| (18) primary copper (*) | (38) network broadcasting (P) |
| (19) integrated copper operations (*) | (39) utilities (P) |
| (20) razor blades and razors (Sh) | (40) precious metals (P) |

□ **Industries with substantial barriers to entry:**

- | | |
|---|---|
| (1) light farm equipment (*) | (7) alkalies and chlorine (MW, Sh) |
| (2) blended and prepared flour mixes (Sh) | (8) industrial gases (Sh) |
| (3) bread and related products (*) | (9) intermediate coal tar products (Sh) |
| (4) chocolate and cocoa products (Sh) | (10) inorganic pigments (Sh) |
| (5) sanitary paper products (Sh) | (11) other inorganic chemicals (Sh) |
| (6) periodicals (*) | (12) organic chemicals (Sh) |
| | (13) synthetic rubber (Sh) |

- | | |
|--|---|
| (14) other man-made fibers (Sh) | (42) household refrigerators (Sh) |
| (15) toilet preparations (Sh) | |
| (16) petroleum refining (B, MR, Sh) | (43) household laundry equipment (Sh) |
| (17) tires (*) | (44) beer (*) |
| (18) other pressed and blown glass (*) | (45) lumber (Sh) |
| (19) cement (*) | (46) fertilizers (Sh) |
| (20) blast furnaces and steel mills (*) | (47) radios and televisions (Sh) |
| (21) cold-finished steel shapes (*) | (48) broadcasting equipment (Sh) |
| (22) steel pipe and tube (*) | (49) electron tubes (Sh) |
| (23) primary zinc (Sh) | (50) cathode ray picture tubes (Sh) |
| (24) primary aluminum (MR, MW, Sh) | (51) storage and primary batteries (Sh) |
| (25) copper rolling and drawing (*) | (52) X-ray apparatus (Sh) |
| (26) aluminum rolling and drawing (Sh) | (53) hard surface floor coverings (*) |
| (27) nonferrous wire drawing, etc. (Sh) | (54) heavy industrial machinery (Sh) |
| (28) metal cans (*) | (55) air conditioning and heating equipment (P) |
| (29) shoe machinery (MR, MW) | (56) plastic production (P) |
| (30) sugar (MW, Sh) | (57) natural gas (P) |
| (31) steel power boilers (Sh) | (58) heavy auto parts (P) |
| (32) safes and vaults (Sh) | (59) paperboard (P) |
| (33) steam engines and turbines (Sh) | (60) local broadcasting (P) |
| (34) internal combustion engines (Sh) | (61) vehicle rental (P) |
| (35) construction machinery (Sh) | (62) natural resources (P) |
| (36) elevators and escalators (Sh) | (63) office machines (P) |
| (37) industrial trucks and tractors (Sh) | (64) insecticides (P) |
| (38) rolling mill machinery (Sh) | (65) steamship operations (P) |
| (39) paper industries machinery (Sh) | (66) lawn and garden equipment (P) |
| (40) typewriters (*) | (67) guns (P) |
| (41) household cooking equipment (Sh) | (68) soup (P) |
| | (69) baby food (P) |
| | (70) computer software (P) |
| | (71) credit cards (P) |
| | (72) refractory materials (P) |
| | (73) mobile homes (P) |
| | (74) cigars (P) |
| | (75) silverware and plated ware (P) |

□ Industries with moderate-to-low barriers to entry:

- | | |
|--|--|
| (1) small metal products (Sh) | (7) flour (B, MR, Sh) |
| (2) wooden furniture (Sh) | (8) footwear (Sh) |
| (3) fluid milk (Sh) | (9) sewing machines (Sh) |
| (4) wet corn milling (Sh, MW) | (10) woolen and cotton textiles (MR, Sh) |
| (5) biscuits, crackers and cookies (*) | (11) clothing (MR, Sh) |
| (6) rayon (*) | (12) glass containers (*) |

- | | |
|---|--|
| (13) brick and structural tile (Sh) | (30) printing (Sh) |
| (14) gypsum products (*) | (31) quarrying (P) |
| (15) abrasive products (Sh) | (32) plastic fabrication (P) |
| (16) electrometallurgical products (Sh) | (33) paperboard products (P) |
| (17) foundries (Sh) | (34) packaging materials (P) |
| (18) secondary non-ferrous metals (M-L) | (35) light auto parts (P) |
| (19) canned fruits and vegetables (B, MR, Sh) | (36) dairy products (P) |
| (20) bituminous coal (MR) | (37) asbestos (P) |
| (21) transportation equipment, hardware (Sh) | (38) real estate (P) |
| (22) metal plumbing fixtures (Sh) | (39) adhesives (P) |
| (23) ball and roller bearings (Sh) | (40) tape and allied products (P) |
| (24) household vacuum cleaners (Sh) | (41) advertising services (P) |
| (25) books (*) | (42) educational and training services (P) |
| (26) corrugated containers (Sh) | (43) protection systems (P) |
| (27) engine electrical equipment (Sh) | (44) leisure products (P) |
| (28) surgical appliances and supplies (Sh) | (45) retailing (P) |
| (29) meatpacking (B, MR, Sh) | (46) crafts, games, and toys (P) |
| | (47) printing ink (P) |
| | (48) restaurants (P) |
| | (49) insurance (P) |
| | (50) finance (P) |
| | (51) rugs and carpets (P) |
| | (52) writing equipment (P) |
| | (53) ophthalmic goods (P) |
| | (54) soft drink bottlers (P) |

References

1. BAIN, J. *Barriers to the New Competition* Cambridge: Harvard Univ. Press, 1965.
2. HALL, M. AND WEISS, L. "Firm Size and Profitability." *Review of Economics and Statistics*, Vol. 49, No. 3 (August 1967), pp. 319-331.
3. JOHNSTON, J. *Econometric Methods*. New York: McGraw-Hill Book Co., 1972.
4. KAMERSCHEN, D. "The Influence of Ownership and Control on Profit Rates." *The American Economic Review*, Vol. 58, No. 3 (June 1968), pp. 432-447.
5. LARNER, R. *Management Control and the Large Corporation*. New York: Dunellen, 1970.
6. LEWELLEN, W. *The Ownership Income of Management*. New York: Columbia Univ. Press, 1971.
7. MANN, H. "Asymmetry, Barriers to Entry, and Rates of Return in Twenty-six Concentrated Industries, 1948 to 1957." *Western Economic Journal*, Vol. 8, No. 1 (March 1970), pp. 86-89.
8. ———. "Seller Concentration, Barriers to Entry, and Rates of Return in Thirty Industries, 1950-1960." *Review of Economics and Statistics*, Vol. 48, No. 3 (August 1966), pp. 296-307.
9. MASSON, R. "Executive Motivations, Earnings, and Consequent Equity Performance." *Journal of Political Economy*, Vol. 79, No. 6 (November-December 1971), pp. 1278-1292.
10. MONSEN, R., CHIU, J., AND COOLEY, D. "The Effect of Separation of Ownership and Control on the Performance of the Large Firm." *Quarterly Journal of Economics*, Vol. 82, No. 3 (August 1968), pp. 435-451.
11. OSTLE, B. *Statistics in Research*. Second ed. Ames, Iowa: The Iowa State University Press, 1963.

12. RHOADES, S. "Concentration, Barriers to Entry, and Rates of Return Revisited." *Journal of Industrial Economics*, Vol. 20, No. 2 (April 1972), pp. 193-195.
13. SHEPHERD, W. *Market Power and Economic Welfare*. New York: Random House, 1970.
14. STIGLER, G. "Imperfections in the Capital Market." *Journal of Political Economy*, Vol. 75, No. 3 (June 1967), pp. 287-292.
15. TELSER, L. "Advertising and Competition." *Journal of Political Economy*, Vol. 72, No. 6 (December 1967), pp. 537-562.
16. U. S. FEDERAL TRADE COMMISSION. "Comparable Concentration Ratios for 213 Manufacturing Industries." Washington, D. C., 1967.

A simple model of firm behavior under regulation and uncertainty

Stewart C. Myers

Associate Professor of Finance
Sloan School of Management

This paper is an analysis of how a firm behaves in the face of uncertainty about demand and cost conditions and a known constraint on the rate of return it is allowed to earn. Under these conditions it is improbable that regulation could force a monopoly to make competitive investment and output decisions. This discouraging result is due to uncertainty per se and not to the phenomena discussed by Averch and Johnson.

1. Introduction

■ This paper presents a theoretical description of the effects of regulation in an uncertain world. The analysis is based on a simple model of how firms behave in the face of uncertainty and a known constraint on the rate of return they can earn.

The main result is an “improbability theorem.” It is not likely that a rate-of-return constraint could force a monopolist to make competitive investment and output decisions. The competitive solution follows only if:

- (1) Special demand and cost conditions obtain at the competitive outputs.
- (2) Regulators are willing to set the allowed rate of return at the *highest* level that might be achieved under competitive conditions.
- (3) The utility does not “cheat,” even when there is ample opportunity and incentive to do so, and little chance of being “caught” by the regulators.

The improbability theorem has nothing to do with Averch-Johnson effects,¹ since the model on which it is based does not consider the firm’s choice of factor proportions.

The paper’s line of attack is (1) to calculate the equilibrium in-

The ideas in this paper have been presented at three conferences. The first was sponsored by the Inter-University Committee on Public Utility Economics and held at Michigan State University in February 1971. The second and third were the A.T.&T. Conferences on Financial Aspects of Utility Regulation, held at Stanford University in June 1971 and June 1972. My audiences at these conferences supplied many helpful comments.

Some of the ideas presented in this paper also appear in “On Public Utility Regulation under Uncertainty” [4], the paper which I presented to the Michigan State conference. That paper includes the rudiments of the model presented here, but it does not include a thorough analysis of the model and its implications. I am grateful to the Bureau of Business and Economic Research at Michigan State for permission to reproduce certain parts of the earlier article here.

Mr. Stavros Thomadakis deserves special thanks for his detailed comments and advice on the paper.

¹ See [1].

vestment and output decisions of a monopolistic firm acting under uncertainty, (2) to compare this equilibrium result with the one that would be achieved under competitive conditions, and (3) to show how regulatory constraints can affect the behavior of such a firm. The analysis is conducted for an extremely simple case with the help of a geometric presentation and a numerical example. I believe this case captures the essential elements of the problem, however.

■ The analysis will employ the “time-state-preference” framework for analysis of behavior under uncertainty.² That is, uncertainty about conditions in a future period is described by specifying a set of possible “states of nature,” only one of which can actually occur. The realized values of random variables are contingent on which state of nature occurs, and we can describe uncertainty about any relevant variable (cost, demand, profit, production capacity, etc.) by listing its possible contingent values.

I will analyze a one-period world, in which there are only two possible future states of nature ($s = 1, 2$). Decisions are made now, at $t = 0$, without knowing which state will occur at $t = 1$. (The generalization to many states is trivial; the generalization to many time periods is not.)

Under ideal conditions, markets will arise for payments contingent on the occurrence of particular states of nature. We will assume that there are no restrictions on trading in these contingent contracts. Thus, the possibility of arbitrage insures that the present market value of any asset is given by

$$\begin{aligned} [\text{Present Market Value}] = & V(1)[\text{Cash payoff if } s = 1] \\ & + V(2)[\text{Cash payoff if } s = 2], \end{aligned}$$

where $V(s)$ is the market value of a contract offering \$1 at $t = 1$ if and only if state s occurs.

■ Consider a firm which has to invest now ($t = 0$) in order to build capacity for production in the future ($t = 1$). Cost and demand conditions for $t = 1$ are not known. They depend on which of the two possible states of nature occurs.

We will employ the following notation, where s indexes the possible states:

$V(s)$ = present value of \$1 to be delivered if and only if state s occurs,

$Q(s)$ = the firm’s output in s ,

Q_M = maximum output, determined by investment at $t = 0$,

$C(s)$ = out-of-pocket cost per unit of output in s ,

$P(s)$ = price per unit of output in s , and

$F(Q_M)$ = investment required to build the capacity Q_M .

The firm’s objective is to maximize its net present market value, subject to the constraint that, regardless of the state of nature oc-

2. Framework for the analysis

3. The basic model

² For a detailed exposition see Hirshleifer, [2].

curing, it cannot produce more than its capacity. Formally, its problem³ is to

$$\text{Max}_{Q(s), Q_M} \psi = \sum_{s=1,2} V(s)[P(s) - C(s)]Q(s) - F(Q_M) \quad (1)$$

subject to $\phi(s) = Q(s) - Q_M \leq 0$. It is interesting to note the similarity between (1) and the peak-load pricing problem.

Assuming $Q(1)$, $Q(2)$, and Q_M all are positive, the necessary conditions for the maximum are:

$$\begin{aligned} Q(1): \quad & V(1) \left[P(1) + Q(1) \frac{\delta P(1)}{\delta Q(1)} - C(1) - Q(1) \frac{\delta C(1)}{\delta Q(1)} \right] \\ & - \lambda(1) = 0 \\ Q(2): \quad & V(2) \left[P(2) + Q(2) \frac{\delta P(2)}{\delta Q(2)} - C(2) - Q(2) \frac{\delta C(2)}{\delta Q(2)} \right] \\ & - \lambda(2) = 0 \\ Q_M: \quad & - \frac{\delta F}{\delta Q_M} + \lambda(1) + \lambda(2) = 0 \end{aligned} \quad (2)$$

$$\text{Also: } \phi(s)\lambda(s) = 0, \quad s = 1, 2.$$

The conditions are easy to interpret. The shadow prices $\lambda(1)$ and $\lambda(2)$ represent the present value of the difference between marginal revenue and marginal cost. Marginal revenue and marginal cost are not equal unless output is less than capacity, in which case $\phi(s) < 0$ and $\lambda(s) = 0$. But the condition for Q_M assures that capacity will limit output in at least one state, since $\delta F / \delta Q_M > 0$. Thus the sum of the shadow prices $\lambda(s)$ also represents the present value of extra capacity, and the condition on Q_M simply states that capacity should be expanded until the marginal cost of an extra unit equals the present value of an extra unit available for use in the various states.

The solution for competitive markets is different in two respects. First, from the viewpoint of the competitive firm, $Q(s) \frac{\delta P(s)}{\delta Q(s)} = 0$ for all s . That is, $P(s)$ appears instead of marginal revenue in the conditions on $Q(s)$. Second, there will be entry into, or exit from, the industry until $\psi = 0$.

There is no reason to expect $\lambda(1)$ to equal $\lambda(2)$ in either the competitive or monopolistic case. This is important, because the λ 's reflect the profitability that the utility will enjoy after the true state of nature is revealed—in other words, they reflect *ex post* profitability. Even in competitive equilibrium, firms may enjoy “monopolistic profits” or incur substantial losses after the fact. Clearly, the attempt by regulators to impose a “reasonable” or “fair” rate of return after the fact may rule out any chance of approximating the competitive solution.

³ A model similar to this one was arrived at independently by S. C. Littlechild in [3]. However, Littlechild does not investigate the effects of regulation.

■ It is a bit artificial to think of a competitive solution in a regulated industry, in which competition is almost by definition imperfect. Nevertheless, regulation has been conceived of as a substitute for competition—that is, as an attempt to enjoy the welfare-maximizing properties of competitive equilibrium despite the existence of a naturally monopolistic or oligopolistic industry.

Thus in the present context ideal regulation would meet two conditions:

Condition 1: Regulation would force the utility to act as if marginal revenue in s equalled $P(s)$ rather than $P(s) + Q(s) \frac{\delta P(s)}{\delta Q(s)}$.

Condition 2: If satisfaction of Condition 1 leads the firm to a solution where ψ is not zero, then the regulators would impose a lump sum tax or subsidy⁴ equal to $-\psi$.

There are no theoretical problems in achieving the second condition, so it will not be discussed further in this paper. Regulation will be said to lead to the competitive solution if the first condition is satisfied.

■ There is a variety of regulatory strategies that might lead to a competitive solution. The impact of these strategies can be analyzed by adding constraints and observing the changes in $Q(1)$, $Q(2)$ and Q_M .

In real life the constraint is on book rate of return—i.e., on the ratio of accounting income to the book value of past investment. Presumably a utility's investment decision and operating plans take account of the fact that regulators will force it to lower the price of its product if its *ex post* rate of return would otherwise exceed some maximum R^* .

Of course the firm does not know what the future value of R^* will be, although an established utility probably can make a good guess. Nor does it know how promptly and effectively the constraint will be enforced. However, I will ignore these complications here. R^* will be assumed known *ex ante*, and it will be assumed that the constraint is strictly enforced.

Thus, the constraints change to

$$\phi_A(s) = Q(s) - Q_M \leq 0$$

and

$$\phi_B(s) = Q(s)[P(s) - C(s)] - (1 + R^*)F(Q_M) \leq 0.$$

(1a)

To simplify notation, let $MP(s)$, "marginal profit" in s , be given by

$$MP(s) = P(s) + Q(s) \frac{\delta P(s)}{\delta Q(s)} - C(s) - Q(s) \frac{\delta C(s)}{\delta Q(s)}.$$

Then the addition of the constraints $\phi_B(s)$ changes the conditions for

4. Regulation as a substitute for competition

5. Rate-of-return regulation

⁴ This assumes, of course, that any required subsidy is less than the consumer surplus generated at the competitive solution.

the maximum to:

$$Q(s): V(s)MP(s) - \lambda_A(s) - \lambda_B(s)[P(s) - C(s)] = 0, \quad s = 1, 2$$

$$Q_M: -\frac{\delta F}{\delta Q_M} + \lambda_A(1) + \lambda_A(2) + \frac{\delta F}{\delta Q_M} (1 + R^*)[\lambda_B(1) + \lambda_B(2)] = 0 \quad (2a)$$

Also: $\phi_A(s)\lambda_A(s) = 0, \quad \phi_B(s)\lambda_B(s) = 0, \quad s = 1, 2.$

A variety of other regulatory strategies can also be analyzed by changing the constraints $\phi_B(s)$, but (1a) seems to be the best simple representation of actual practice.

6. A numerical example

■ The implications of (2a) turn out to be surprisingly complex and difficult to interpret. Therefore, it will help to start with a numerical example. We assume that the firm is a monopoly, and faces the following contingent demand functions:

$$Q(1) = 2200 - 400P(1)$$

$$Q(2) = 3000 - 400P(2),$$

thus

$$P(1) = 5.5 - 0.0025Q(1)$$

$$P(2) = 7.5 - 0.0025Q(2).$$

Costs are also contingent on the state occurring: $C(1) = 1.0$ and $C(2) = 1.5$. The present investment required per unit of capacity is 1.0; i.e., $F(Q_M) = (1.0)Q_M$. Present values are obtained via the assumed prices $V(1) = 0.5$ and $V(2) = 0.4$. This implies a risk-free rate of interest of about 11 percent.

Using the conditions stated in (2a), the optimal solution for $Q(1)$, $Q(2)$ and Q_M can be calculated. The values are shown in the first column of Table 1, along with several other characteristics of the solution. These may be compared with column two of the table,

TABLE 1
RESULTS FOR NUMERICAL EXAMPLE

	MONOPOLY SOLUTION	COMPETITIVE SOLUTION	CONSTRAINT ON MAXIMUM RATE OF RETURN ($R^* = 0.20$)	
			RESPONSE A	RESPONSE B
$Q(1)$	812	1622	900	1320
$Q(2)$	812	1622	1920	1320
Q_M	812	1622	1920	1320
REQ'D INVESTMENT	812	1622	1920	1320
NET PRESENT VALUE	1486	0	+12 $\approx$ 0	106
$R(1)$, EX POST RATE OF RETURN IN $s = 1$	1.47	-0.555	0.05	0.20
$R(2)$, EX POST RATE OF RETURN IN $s = 2$	2.97	0.945	0.20	0.20
$P(1)$	3.47	1.445	3.25	2.2
$P(2)$	5.47	3.445	2.7	2.2
$\frac{1 + R(2)}{1 + R(1)}$	1.61	4.37	1.14	1.0

which shows the solution if the monopoly could be forced to behave like a perfect competitor. The results for the competitive case were calculated from (2), but with price substituted for marginal revenue in each state.

The differences between the monopoly and competitive solutions are clear from the first two columns of Table 1. Only a few comments are called for. When the firm is free to act as a monopoly, it restricts capacity and output, charges high prices, and earns a whopping profit. The profit is reflected in a high positive net present value. In the competitive case there is no such *ex ante* windfall gain,⁵ although profits are high *ex post* if state 2 occurs.

In this two-state model, a firm's risk may be simply described by the *ratio* of its payoffs in states 1 and 2—or, alternatively, by the ratio $[1 + R(2)]/[1 + R(1)]$, where $R(s)$ is the *ex post* rate of return in s . (In a state-preference framework, the riskier assets are those which pay off relatively more in states with low V 's and less in states with high V 's.) Thus, we see from Table 1 that the monopoly solution for this example results in a safer firm than the competitive solution. Incidentally, although the monopoly is safer in this instance, this is *not* a general result.

Now let us see how the firm reacts to a regulatory constraint of $R^* = 0.20$. There are a number of responses the firm might make. One would be to expand investment and output in $s = 2$ until $R(2) = 0.20$, and to maximize profits if state 1 occurs, subject of course to the constraints $\phi_A(1)$ and $\phi_B(1)$. The results for this case are shown in the third column, headed "Response A," in Table 1.

Given this response, the effect of the constraint is to drive the net present value of the utility from 1486 to approximately zero, i.e., to eliminate the *ex ante* monopoly profit. This presumably is a good thing. However, other characteristics of the constrained solution are not so desirable. The firm is led to invest more than it would in the competitive case; it does this to drive down the *ex post* rate of return in $s = 2$. This additional capacity is used only in state 2, however. The firm has invested enough so that the constraint on maximum profit is not binding in $s = 1$, and so the firm seeks maximum profit in that state, restricting output and driving up prices. Even at an output of 900, where $MP(1) = 0$, it can only earn 5 percent on the investment of 1920. In short, the regulatory constraint would eliminate *ex post* "monopoly profits" in one state at the expense of allowing the firm freedom from regulation if the other state occurs. It is also interesting that in this example the firm becomes safer from the investor's point of view when it operates under the constraint.

It must be emphasized that this is one of several ways in which the firm might react to imposition of a rate-of-return constraint. But I will consider the other responses later.

■ The consequences of the model just described can be clarified by a geometric presentation. I will first present the solution for an unconstrained monopoly and then show how the firm can respond when a constraint is imposed on *ex post* rate of return.

⁵ The numerical example assumes constant returns to scale, which implies that $\psi = 0$ when $[\delta P(s)/\delta Q(s)]Q(s)$ is deleted from (2). Thus, we do not have to introduce a tax or subsidy to obtain the competitive solution.

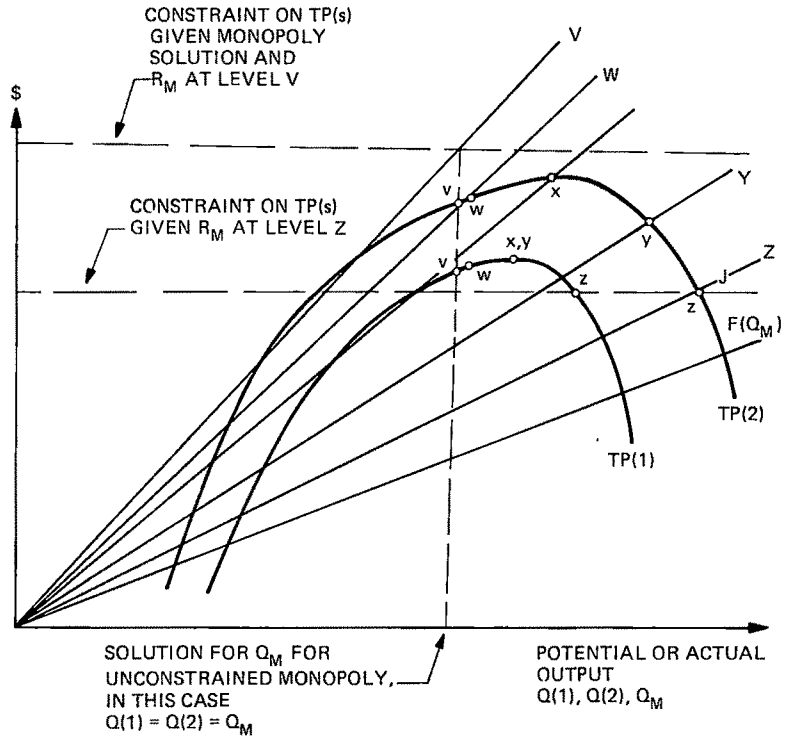
7. A geometric treatment of the model

For convenience, it will be assumed that $\delta F/\delta Q_M = f$, a constant. That is, the marginal investment required to add a unit of capacity is constant. Also, let $TP(s) = Q(s)[P(s) - C(s)]$, the total cash return realized in state s . Note that $MP(s) = \delta TP(s)/\delta Q(s)$.

Figure 1 shows $TP(s)$ for each state as a function of potential or actual output. In the case shown, both $MP(1)$ and $MP(2)$ are positive at the optimal Q_M , so the monopoly produces at capacity regardless of which state occurs. (This would be a frequent result in highly capital intensive industries.)

FIGURE 1

EFFECTS OF EX POST RATE OF RETURN CONSTRAINT ON A UTILITY'S OUTPUT AND INVESTMENT DECISIONS



The lowest solid ray extending from the origin in Figure 1 represents F , the amount of investment as a function of Q_M . Any constraint on *ex post* rate of return limits $TP(1)$ and/or $TP(2)$ to an amount proportional to F . Given R^* , we can thus represent the level of the constraint as a function of Q_M by a ray lying above F . The higher R^* , the higher the slope of the ray.⁶ Given Q_M the rays Z , Y , X , W , and V define a series of constraints resulting from successively higher values of R^* .

Once Q_M is established, the constraint on total profits is independent of output in either state. Thus, if capacity is that called for by the monopoly solution and if R^* is set at level V , then the constraint on $TP(1)$ and $TP(2)$ may be represented by the higher dashed horizontal line.

Now we can see how the utility's output and investment decisions change if R^* is set at a high value (level V) but then gradually lowered. We continue to assume that the utility's response to a bind-

⁶ The constraints shown are linear because F is assumed to be a linear function of Q_M . This is not necessary to the argument.

ing constraint is to lower price and increase output in the binding state.⁷

Level W: At this level the monopoly solution violates the constraint in state 2. The utility is thus forced to increase investment Q_M . Since $MP(1)$ and $MP(2)$ are both positive at the monopoly solution and at the new levels of output and capacity, both output and “total profits” $TP(s)$ increase in each state. The net present value of the firm declines, however.

Level X: As R^* is decreased to level X , the utility is forced to increase Q_M and $Q(2)$ still further to satisfy the constraint in state 2. $Q(1)$ will be increased also, but only up to the point at which $TP(1)$ is maximized. There will be excess capacity in $s = 1$ because the constraint on *ex post* rate of return is not binding in $s = 1$, and because increasing output to $Q(1) = Q_M$ would reduce $TP(1)$. In other words, the utility can act like an unconstrained monopolist in that state.

Level Y: Reducing R^* to level Y does not change $Q(1)$, but forces the utility to increase $Q(2)$ and Q_M . The effect is to decrease $TP(2)$. (Note that the results obtained for Response A in Table 1 are of type X or Y .)

Level Z: Finally, at level Z , the constraint is binding in both periods. As R^* is moved from level Y to Z , both $Q(1)$ and $Q(2)$ are increased, although the utility will still have excess capacity in state 1. Since $TP(1) = TP(2)$, the effect of the constraint at level Z is to make the utility an absolutely safe firm—assuming, of course, that it is still willing to stay in business at this point.

There is more to the matter than that, however. A close look at Figure 1 shows the marginal effect of a small decrease in R^* , starting from each of the levels, V , W , X , Y and Z . In this context, it is easiest to interpret “risk” as the ratio of $TP(2)$ to $TP(1)$. At V there is of course no change in risk. At W , the effect is ambiguous, since both $TP(1)$ and $TP(2)$ are affected. At X risk increases, since $TP(2)$ increases as R^* declines, but $TP(1)$ is unchanged. Finally, risk decreases at levels Y and Z , since $TP(2)$ decreases and $TP(1)$ is either constant or equal to $TP(2)$. In short, the conclusion is that risk may actually increase if R^* decreases from relatively generous levels. Ultimately, however, a decrease in R^* makes the utility safer, if it is in business at all.

■ All this has assumed that the firm responds to the rate-of-return constraint by investing more and producing more in the high-profit state. However, rationing may be better from the firm’s point of view.⁸

Figure 2 is Figure 1 redrawn, except that the effects of the rate-of-return constraint are shown only for level Y . The type of response

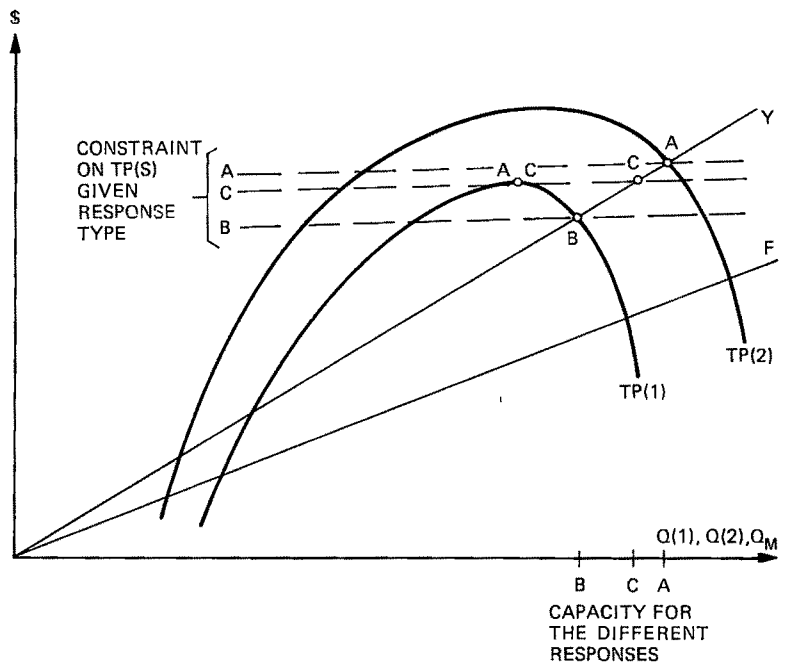
8. Rationing response to the regulatory constraint

⁷ Another alternative is to increase price, reducing demand further and further below capacity until the rate-of-return constraint is finally satisfied. However, this response has less practical interest, since it implies that utilities would raise price and produce less in response to an unexpected upward shift in demand—which is exactly opposite to how actual regulation works.

⁸ Another regulatory strategy is to allow the firm to price so as to clear markets, but to impose a 100-percent tax on any profits in excess of $R^*F(Q_M)$. But that is exactly the same as rationing from the firm’s point of view, so its behavior will be the same as if excess profits were dissipated by rationing rather than the tax.

FIGURE 2

ALTERNATIVE RESPONSES TO RATE OF RETURN CONSTRAINT



discussed so far will lead to the investment and output decisions labelled "A."

Consider "Response B." Here the firm invests and produces only enough to satisfy the rate-of-return constraint in state 1, the "low-profitability" state. Given this lower level of investment the firm could earn substantially more than the minimum return R^* in $s = 2$. However, it keeps price low enough to satisfy the rate-of-return constraint. The excess demand is cured by rationing.

Now turn to the fourth column of Table 1 which shows the results for the numerical example, given Response B. Compared to Response A, we find that although investment is lower, the firm's net present value is increased. The increase occurs because the firm now earns the maximum 20-percent return in each state rather than in state 2 only; the present value of the decrease in $TP(1)$ and $TP(2)$ is more than offset by the reduced investment.

Actually, the firm still has a better option. It could increase capacity to C on Figure 2, thereby allowing it to produce and earn more in each state. Response C is preferable to B so long as R^* exceeds the risk-free interest rate.

It does not follow that rationing will always be a desirable strategy for the constrained firm. Nevertheless, planned rationing for some states of nature strikes me as a very reasonable strategy for a utility. First, the alternative of satisfying demand in all states requires a larger investment which probably will not earn a satisfactory return in low-demand states. Second, the firm will not be "caught" unless demand turns out to be high, and if it is caught the consequences are not fatal. Casual observations of actual recent rationing situations (e.g., of telephone service in New York City and elsewhere) suggests the following scenario. The firm announces an "unpredictable" demand increase and undertakes a crash program to meet demand.

The regulators chide the firm for “poor planning” but accept rationing as an inescapable short-run evil. The firm does lose goodwill with regulators and the public, however.

■ The main conclusion of this paper is probably clear by now, but it should nevertheless be stated precisely and proved. Consider a monopolistic firm operating in the idealized one-period world I have assumed. The object of regulation is to find a value for R^* such that the firm will act as if it were in perfectly competitive markets. This “competitive solution” will be denoted by $\hat{Q}_M$, $\hat{Q}(s)$. Assume there is uncertainty about demand and/or cost conditions, and that the firm’s profits are uncertain at the competitive solution $\hat{Q}_M$, $\hat{Q}(s)$.

Is there always *some* value of R^* that would lead to $\hat{Q}(s)$, $\hat{Q}_M$? Unfortunately, the answer is no, even if we rule out rationing as a possible response by the utility.

Suppose the firm has somehow arrived at a provisional investment and output plan which calls for $\hat{Q}(s)$, $\hat{Q}_M$. It clearly will not stick with this plan once it considers the alternatives. However, is there some rate-of-return constraint which will prevent the firm from deserting the competitive solution?

The conditions under which there will be such a constraint are as follows. First, the regulators must rule out rationing. Second, they must set

$$R^* = \hat{R}(z), \quad (3a)$$

where $\hat{R}(z)$ is the highest rate of return that might be achieved in the competitive case. (This means that the rate-of-return constraint will not be binding in the other states.)⁹ The reason for requirement (3a) is simple. If $R^* > \hat{R}(z)$, then the firm can decrease $Q(z)$ from $\hat{Q}(z)$. If $R^* < \hat{R}(z)$, then the firm is forced to increase $Q(z)$ and probably Q_M above $\hat{Q}(z)$ and $\hat{Q}_M$.

Third, the firm must be operating at capacity in state z :

$$\hat{Q}(z) = \hat{Q}_M, \text{ i.e., } \hat{P}(z) \geq \hat{M}C(z). \quad (3b)$$

Otherwise there is an opportunity for the firm to decrease Q_M .

Fourth, we have to consider the other states $s \neq z$, in which $\hat{R}(s) < R^*$. The firm must have no incentive nor opportunity to change price or output in these states. This will occur if, for all $s \neq z$,

$$MR(s) > MC(s) \text{ at } Q(s) = \hat{Q}_M, \quad (3c)$$

which also implies that a competitive firm would operate at capacity in all states.

These conditions allow regulators to control Q_M by forcing $R(z)$ and thus $Q(z)$ to be competitive levels. Once capacity is given at Q_M , it is in the monopolist’s own interest to adopt the competitive solution $\hat{Q}(s) = \hat{Q}_M$ for the states $s \neq z$.

It does not seem likely that all of the conditions hold, although admittedly each will be more likely to hold, the more capital-intensive the firm is. However, this is not a very comfortable result even if the

9. An “improbability theorem”

⁹ This assumes that the maximum competitive rate of return occurs in only one state, i.e., that $\hat{R}(s) < \hat{R}(z)$ for all $s \neq z$. It is conceivable that there will be some state y in which $\hat{R}(y) = \hat{R}(z)$. However, the discussion in the text can be applied to y as well as to z .

conditions (3b) and (3c) really are true. Since R^* would be set at the highest possible competitive rate of return, the rate-of-return constraint would be binding *ex post* only a small percentage of the time. A utility could violate it *ex ante* with confidence that it would be discovered only rarely.¹⁰ If the regulatory authorities reduced R^* to “catch” the utility more frequently, it would drive the utility towards the type of behavior illustrated by Responses *A*, *B*, and *C* in Table 1 and Figure 2.

Thus, we have an “improbability theorem”—a fancy label for saying that it seems unlikely that a straightforward rate-of-return constraint could be used to enforce competitive behavior under uncertainty. Of course, this result is obtained under idealized assumptions, but it is unclear what factors could allow rate-of-return regulation to perform better in real life.

Note that the improbability theorem is due specifically to uncertainty. The conditions just stated are not stringent if there is only one state of nature. Note also that the improbability theorem has nothing to do with Averch-Johnson effects, since the model presented in this paper does not consider the firm’s choice of factor proportions.¹¹

10. Summary and conclusions

■ This paper presents a theory of regulation under uncertainty, based on a very simple model of firm behavior. The main result is an “improbability theorem,” which states that rate-of-return regulation can force a monopolist to make competitive investment and output decisions only under very special conditions, namely:

- (1) Special requirements on demand and cost conditions at the competitive outputs.
- (2) The willingness of regulators to set the allowed rate of return equal to the highest rate of return that might be achieved under competitive decisions.
- (3) The willingness of utilities to forego “cheating” under the loose regulation required by item 2.

However, this paper is not a plea for tossing rate-of-return regulation into the ash can right away. First, there is no evidently superior alternative. Second, we lack an empirical assessment of the extent of departure from the competitive solution. Third, rate-of-return regulation can at least provide equitable treatment of consumers and utility investors.

¹⁰ The cheating would take the form of planned rationing in z and other high-demand states. Let us assume conditions (3a), (3b) and (3c) hold. We know that if the constraint on R^* were loosened, a monopoly would increase its net present value by decreasing $Q(s)$ and Q_M from the competitive solution. A reduction in $F(Q_M)$ would be less than offset by the reduction in the present value of marginal profits.

If the constraint on R^* were not loosened, but the firm nevertheless reduced $Q(s)$ and Q_M , the reduction in the present value of marginal profits would be greater because of rationing in state z . However, it seems very likely that the firm would still gain, since rationing affects profits in only one state.

¹¹ However, when this choice is included in the model, a result closely analogous to Averch and Johnson’s is obtained. Specifically, it can be shown that a regulated firm facing uncertain demand and cost conditions will choose a more capital intensive technology than is efficient, given its chosen capacity and outputs.

References

1. AVERCH, H. AND JOHNSON, L. L. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol. 52, No. 5 (December 1969), pp. 1053-1069.
2. HIRSHLEIFER, J. *Investment, Interest and Capital*. Englewood Cliffs, N.J.: Prentice-Hall, 1970.
3. LITTLECHILD, S. C. "A State Preference Approach to Public Utility Pricing under Risk." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 1 (Spring 1972), pp. 340-345.
4. MYERS, S. C. "On Public Utility Regulation under Uncertainty," in R. H. Howard, ed., *Risk and Regulated Firms*, East Lansing, Mich.: Division of Research, Graduate School of Business Administration, Michigan State University, 1973.

Pricing of pollution: the Coase Theorem in the long run

H. E. Frech III

University of California, Los Angeles
The National Center for Health Services Research and Development, DHEW

In an earlier article in The Bell Journal, Tybout argues that even in a zero transaction costs model, bribery to reduce pollution and compensation charges for it result in different total profits, and thus in different long-run behavior. Therefore, the Coase Theorem is refuted for the long-run case.

This paper demonstrates that Tybout's result is based on a fundamental error in the analysis of costs for the polluting and polluted industries. The property right to the polluted (air or water) basin is a valuable asset with a definite market value. When the rent of this valuable asset is properly included in the costs of the industry holding the right, marginal and total profits are shown to be identical regardless of the assignment of the basin ownership.

1. Introduction

■ In an earlier article in this *Journal*, Tybout attempted a refutation of the long run applicability of the Coase Theorem.¹ Leaving aside distributional effects on demand and assuming zero transaction costs, Coase asserted that resource allocation is identical whether pollution externalities are internalized by the polluter paying charges (compensation) or by the recipient paying the polluter to withhold pollution (bribery).²

The refutation hinges on Tybout's finding that total profits for each of two industries (one which pollutes and the other which is hindered by the pollution) differ according to whether compensation or bribery is charged. And "total profits cannot help influencing the adjustments."³ Bramhall and Mills, Wellisz, and Calabresi⁴ have also maintained that differing total profits will affect exit and entry and thus influence long-run resource allocation. As Wellisz noted, given this profit difference, the Coase Theorem would hold in the long run only if one postulated nontransferable resources' earning rents.

Those defending the Coase Theorem for the long run have included Calabresi, recanting his earlier views, Demsetz, and Nutter.⁵ However, Demsetz gave an example with nontransferable resources (land) earning rents, seemingly leaving open the case of transferable resources. Nutter simply reasons that if the allocation of rights is

Thanks are due to Paul B. Ginsburg and Larry Manheim for helpful discussions. Neither affiliated institution necessarily agrees with the views expressed herein.

¹ In [7].

² In [4].

³ See [7], p. 261.

⁴ In [1], [8], and [2], respectively.

⁵ In [3], [5], and [6], respectively.

inefficient, the firms concerned will merge. While both analyses are formally correct, one would prefer the Coase Theorem to hold even if all resources are transferable or if the merger is foreclosed.

■ Since the essential point of Tybout's argument is that *total* profits to the industries concerned differ according to the assignment of pollution liability, we concentrate on this. If compensation is required, pollution is a bad for which the polluter is charged a market price. Under a regime of compensation where the polluter is liable, the profit of industry *A* (the polluter) and of industry *B* (the recipient) is given by:⁶

$$\pi_A = P_A X_A - KS - wL_A - rC_A = 0 \quad (1a)$$

$$\pi_B = P_B X_B + KS - wL_B - rC_B = 0, \quad (3a)$$

where *P* ≡ Price, *X* ≡ Quantity, *w* ≡ Wage, *L* ≡ Quantity of Labor, *r* ≡ Return on Capital, *C* ≡ Quantity of Capital, *K* ≡ Charge per Unit of Pollution, and *S* ≡ Equilibrium, Quantity of Pollution. The equilibrium quantity of pollution, *S*, is determined by supply and demand analysis where the marginal profits of pollution to industry *A* are equated to the marginal losses to industry *B* at the market price of the pollution, *K*.

If industry *B* must bribe *A* to reduce pollution, the good in question becomes the withholding of pollution. If the amount of pollution which would be produced in the absence of any bargains is *Z*, then the profits for the two industries in a regime of bribery are given by:⁷

$$\pi_A = P_A X_A + K(Z - S) - wL_A - rC_A = KZ \quad (1b)$$

$$\pi_B = P_B X_B - K(Z - S) - wL_B - rC_B = -KZ. \quad (3b)$$

Clearly, if these formulations are correct, total profits differ between the two liability rules. If resources are transferable, long-run resource allocation will be affected. Thus, Tybout maintains, the Coase Theorem is refuted for the long run.

Aside from overlooking the possibility of integrating the two industries, Tybout's fundamental error is omitting the asset value (and rent thereof) of the right to control pollution of the environment. This property right to the use of the (air or water) basin is a valuable resource. Only by ignoring the return to that resource does Tybout obtain his result that total profit differs according to the assignment of the rights concerned.

■ Within Tybout's competitive pricing model, the value of ownership rights to the polluted basin for industry *A* is the present value of the bribes received, *K*(*Z* - *S*) plus the compensation charges avoided, *KS*. For industry *B*, the value of the asset is the present value of the bribes avoided, *K*(*Z* - *S*) plus the compensation received, *KS*. The

2. Tybout's analysis

3. The value of the property right to the polluted basin

⁶ See [7], p. 258.

⁷ *Ibid.*, p. 259.

total, regardless of ownership, is the present value of KZ .⁸ Thus $r\left(\frac{KZ}{r}\right)$, is the imputed rent to basin ownership.

4. Profits including the value of the basin ownership

■ If the rights to the basin are assigned to industry B (compensation), the profits of the two industries are given by:

$$\begin{aligned}\pi_A &= P_A X_A - KS - wL_A - rC_A = 0, \\ \pi_B &= P_B X_B + KS - wL_B - rC_B - r\left(\frac{KZ}{r}\right) = 0.\end{aligned}\tag{1a}$$

Rearranging terms gives:

$$\pi_B = P_B X_B - K(Z - S) - wL_B - rC_B = 0.\tag{3a'}$$

If the property rights are granted to industry A (bribery), the profits to the two industries are given by:

$$\pi_A = P_A X_A + K(Z - S) - wL_A - rC_A - r\left(\frac{KZ}{r}\right) = 0.$$

Rearranging terms gives:

$$\pi_A = P_A X_A - KS - wL_A - rC_A = 0\tag{1a}$$

$$\pi_B = P_B X_B - K(Z - S) - wL_B - rC_B = 0.\tag{3a'}$$

When the imputed rent of the asset of ownership rights to the pollution basin is properly accounted for, total profits are identical regardless of ownership. Therefore, behavior is identical. The Coase Theorem holds.

A word on the economic interpretation of the expenditure quantities, KS and $K(Z - S)$ may be worthwhile. For the polluting industry, pollution is a factor of production. KS is the marginal product of the factor, K , times the quantity used, S . For the recipient industry, withholding pollution from the independent adjustment level, Z , is a factor of production. $K(Z - S)$ is the marginal productivity of withholding, K , times the quantity consumed ($Z - S$). Linear homogeneity, assumed by Tybout, requires that profits be zero in equilibrium (product exhaustion).

5. Conclusion

■ Abstracting from distributional effects on demand and transaction costs, as do all participants in the debate, resource allocation is unaffected whether the property rights to the polluted basin are assigned to the recipients (compensation), to the polluters (bribery), or to a third party who rents the asset. The initial assignment of the rights concerned does affect the distribution of wealth as a purely *windfall* gain or loss.

⁸ In either case the value of the resource includes explicit receipts and implicit savings.

References

1. BRAMHALL, D. F. AND MILLS, E. S. "A Note on the Asymmetry between Fees and Bribes." *Water Resources Research*, Vol. 2, No. 3 (July-September 1966), pp. 615-616.

2. CALABRESI, G. "The Decision for Accidents: An Approach to Nonfault Allocation of Costs." *Harvard Law Review*, Vol. 78, (1965), pp. 713-731.
3. ———. "Transaction Costs, Resource Allocation and Liability Rules—A Comment." *Journal of Law and Economics*, Vol. 11, No. 1 (April 1968), pp. 67-74.
4. COASE, R. H. "The Problem of Social Costs." *Journal of Law and Economics*, Vol. 3, No. 2 (October 1960), pp. 1-44.
5. DEMSETZ, H. "When Does the Rule of Liability Matter?" *Journal of Legal Studies*, Vol. 1, No. 1 (January 1972), pp. 13-28.
6. NUTTER, G. W. "The Coase Theorem on Social Cost: A Footnote." *Journal of Law and Economics*, Vol. 11, No. 2 (October 1968), pp. 503-507.
7. TYBOUT, R. A. "Pricing Pollution and Other Negative Externalities". *The Bell Journal of Economics and Management Science*, Vol. 3, No. 1 (Spring 1972), pp. 252-266.
8. WELLISZ, S. "On External Diseconomies and the Government-Assisted Invisible Hand." *Economica*, Vol. 31, No. 4 (November 1964), pp. 345-362.

Pricing of pollution: reply

Richard A. Tybout

Professor of Economics
The Ohio State University

The conditions of internalization to which this note refers were derived by a general equilibrium analysis independent of site rentals, and would not be changed by properly including site rentals.

■ The property rights with which we should begin are to either nectar or pollution, the latter by enforced ownership. They are not to a "polluted basin." If there are any special gains to site location, such gains should be derived from income streams resulting from the commodity rights.

The opportunity costs of the commodity nectar or the discommodity pollution are found in the uses of resources precluded by the production of either product. These opportunity costs are exactly compensated in a general equilibrium framework, at least with internalization as normally conceived for nectar and with compensation for pollution. The entire amount of the compensation is used in this way, leaving nothing to be imputed to any other property rights and resulting in equations (1a) and (3a) as shown in Tybout.¹ This means that the term $r \left(\frac{KZ}{r} \right)$ in Mr. Frech's first equation (1a) equals zero.²

In order to make a property right purely a windfall, it is necessary that the allocation of resources not be affected by the way that right is assigned. Wellisz' analysis on this point has been noted.³ With transferable resources, however, industry *B* can locate elsewhere and industry *A* can take its pollution elsewhere. Either way, there is a change in the allocation of resources.

One might argue that even though *B* has no excess returns with compensation, *A* does with bribery. True enough, as indicated by equation (1b) and (3b),⁴ but there is no limit on entry to industry *A* and no charge for admission. There may be site value effects on land that can be used by *A* to victimize *B*, but the same general non-existence of an equilibrium value for profits in *A* will preclude an equilibrium value for use of the affected sites.

Site rentals and agglomeration effects can be introduced in the compensation case for pollution production and bear the same relation to general equilibrium analysis as in conventional commodity (say, widget) production. More specifically, the introduction of spatial considerations would be reflected in site rentals due to transportation costs in the case of widgets and site rentals due to environ-

¹ In [2], p. 258. See Section 2, "Economics of Internalization."

² In [1].

³ In [2], p. 259.

⁴ *Ibid.*

mental dilution in the case of pollution, but would in no way change the principles found in the Spring 1972, *Bell Journal* article to which this note refers.

References

1. FRECH, H. E., III. "Pricing of Pollution: The Coase Theorem in the Long Run." *The Bell Journal of Economics and Management Science*, Vol. 4, No. 1 (Spring 1973), pp. 316-319.
2. TYBOUT, R. A. "Pricing Pollution and Other Negative Externalities." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 1 (Spring 1972), pp. 252-266.

Regulated utilities and equipment manufacturers' conspiracies in the electrical power industry

E. David Emery

Assistant Professor
St Olaf College

The paper demonstrates that only with some form of reproduction cost valuation for rate base capital is it possible that the regulatory process might encourage electrical utilities to cooperate or acquiesce with a manufacturers' conspiracy to increase equipment prices. Based on the process function for steam-electric generation developed in the paper, a monopolistic utility subject to rate-of-return regulation using original cost valuation will always have a more profitable alternative than paying conspiratorially inflated equipment prices. Empirically, the acquisition prices for steam-electric generating units purchased by utilities were higher during a "conspiracy" period, 1956–1959 than during the adjacent years, 1954–1955, 1960–1961. Although the statistical results were somewhat ambiguous, it could not be concluded that the type of rate-base valuation scheme applicable to each utility made a significant difference either in the level of prices or in the increase in prices for similar generating units.

1. Introduction

■ Following the "great conspiracy" in the electrical equipment manufacturing industry, some observers noted reluctance on the part of the injured parties to seek redress under Section Seven of the Sherman Act. Subsequently, F. Westfield presented a model demonstrating the possibility that a firm subject to rate-of-return regulation could increase profits with an increase in the acquisition price paid for capital inputs. Westfield concluded:¹

. . . (i) that it can be in the interest of the regulated private power generating company to pay a higher rather than a lower price for the plant and equipment it purchases, and (ii) that in other instances, though not benefited, the regulated utility is not at all harmed by the inflated prices that result from conspiracies.

On the basis of the model, Westfield then cited the regulatory process as an accessory to the conspiracy:²

For the reader on search for a villain to share the stage with the electrical-equipment manufacturers, I suggest it is not that electrical utility which may have had benefit from the conspiracies, but the regulatory process itself.

While Westfield's first conclusion is theoretically defensible, it has limited empirical applicability, and the implications for regulatory policy were not apparent in the broad condemnation of the regulatory process. The role of rate-of-return regulation in encouraging con-

The author is indebted to Frank Gery, Tom Giddings, and the editors of *The Bell Journal* for comments.

¹ In [14], p. 424.

² *Ibid.*, p. 442.

spiracies among utility suppliers depends upon the properties of the production function for the regulated good and upon the method of rate-base valuation used by the regulatory agency. Consequently, it cannot be concluded that the regulatory process will *necessarily* encourage equipment price conspiracies.

In Section 2 of the paper, we demonstrate that with actual cost valuation schemes for rate-base capital, the conditions necessary to induce steam-electric utilities to be willing participants in an equipment price conspiracy are not met. Only where some form of reproduction cost or "fair value" rate-base valuation of capital exists is there a theoretical basis to cite the regulatory process as a possible accessory to an equipment conspiracy. And only in those cases might we expect to find utilities acquiescing to conspiratorially higher equipment prices.

In Section 3, we examine differences in acquisition prices for large steam-electric generating units purchased by utilities subject to differing types of rate-base valuation. For the time period, 1954 to 1961, it cannot be concluded that the acquisition prices paid by utilities subject to some form of reproduction cost valuation were systematically higher than the prices paid by utilities subject to some form of actual cost valuation for similar equipment. During a period of effective "collusion," 1956 to 1959, there was a very significant increase in acquisition prices paid by utilities subject to reproduction cost valuation, while the price increase to utilities subject to actual cost valuation was not significant. However, the *difference* between the increases was not statistically significant.

■ Rate-of-return regulation is a rate-making technique which bases the price of the regulated product upon the projected rate of earnings on assets devoted to production in the regulated industry, i.e., the rate base.³ Regulatory agencies are presumed to set the price of electricity (the weighted general level of the tariff) such that the expected total revenue minus expected operating expenses for the coming period will yield no more than the allowed return on invested capital. The initial Averch-Johnson model and subsequent papers have demonstrated that a regulated firm may regard rate-of-return regulation as imposing a constraint upon earnings and may behave in a manner different from that of an unregulated firm. However, different forms of earnings constraints will provoke different reactions.

□ **Rate-base valuation.** There are three methods commonly used by regulatory agencies in the United States to arrive at a value for the capital equipment devoted to production by a utility: actual cost, reproduction cost, and "fair value." With all three methods only those items judged prudent and necessary are counted in the rate base, and accumulated depreciation is subtracted from the adjudged value, however determined, in order to arrive at a value for rate-base purposes. Actual cost valuation methods assign a value to the capital equipment at some measure of the original cost of the goods when first devoted to public service. Variations in the price of similar

2. Actual cost versus reproduction cost regimes

³ The rate base is defined as the "value" of the capital equipment used by the utility for production in the regulated industry. Other definitions are possible [1].

equipment at a later date would not influence the valuation of equipment already in the rate base. Reproduction cost valuation methods attempt to fix the value of the capital equipment at the cost required to duplicate the system today. With reproduction cost methods, the rate-base value of existing capital equipment depends in part upon the current prices paid by the utility for new equipment. Finally, the so-called fair-value method represents a combination of the actual and reproduction cost methods. In arriving at a value for capital goods, the regulatory agency must consider the actual cost of the equipment and the cost of reproducing the system, and then arrive at some subjectively weighted average of the two. The fair-value scheme is grouped with the reproduction cost scheme since the rate-base value of existing capital goods is again a function of current prices for new equipment.

Rate-base value is a function of the capital goods (K) devoted to production in the regulated industry. The unit of measurement for the capital input (K) used here is the dollar value of the capital equipment at competitive market prices, i.e., the average cost of producing the equipment. But as demonstrated by Westfield, under certain conditions the utilities have an incentive to pay a premium above competitive prices for new equipment. The acquisition price of capital equipment will then consist of a pure or competitive price multiplied by some factor (λ) to reflect any premium paid by the acquiring utility. The acquisition price of a capital good purchased in a competitive market will just be k dollars, ($K = \sum k_i$). Non-competitive behavior on the part of the equipment suppliers or negotiated higher equipment prices by the utility will cause the acquisition price paid by the utility ($k \cdot \lambda$) to exceed the competitive price for capital equipment and λ will have a value greater than one. The rate-base value of capital equipment depends on both the pricing scheme under which the capital goods were acquired and the valuation scheme applied by the regulatory body. (See Table 1.)

TABLE 1
RATE-BASE VALUES FOR
CAPITAL GOODS

PRICING SCHEME	VALUATION SCHEME	
	ACTUAL COST	REPRO- DUCTION COST
COMPETITIVE PRICES	$K \cdot 1$	$B(K, 1)$
PRICE PREMIUM ($\lambda > 1$)	$K \cdot \lambda$	$B(K, \lambda)$

□ **Adjustment to an earnings constraint.** Consider a simplified model of a regulated, franchised monopolistic firm similar to the models proposed by Averch and Johnson, Westfield, Takayama,⁴ *et al.* The firm is assumed to be a profit maximizer, producing a single homogeneous product (Z), using two inputs: a variable input (V) and capital (K). The total revenue function for the firm may take any of the forms:

$$R(Z) = R[f(K, V)] = R(K, V). \quad (1)$$

We make the usual assumptions concerning the properties of the revenue function and assume that taxes and depreciation are both equal to zero which simplifies without changing the essential results of the model. The input prices (W_1 for capital inputs on an annual basis and W_2 for variable inputs) are viewed by the firm as parameters. The allowed rate of return, γ , is set by the regulatory agency at some constant percentage of the rate base and γ is also viewed by the firm as a parameter. The allowed rate of return γ is also assumed to exceed the cost of capital, W_2 .

Assuming that the price premium factor (λ) is a constant for all capital goods purchased, a typical rate-of-return constraint for a

⁴ In [1], [14], and [9], respectively.

utility can be written:

$$\frac{R(K, V) - W_2 \cdot V}{B(K, \lambda)} \leq \gamma. \quad (2)$$

The constraint is assumed to be effective.⁵

The constraint can be reformulated to provide some insight into the basic nature of the argument concerning acquisition of capital goods at premiums above competitive prices. Assuming actual cost valuation, the earnings constraint can be written:

$$R(K, V) - W_2 \cdot V \leq \gamma \cdot K. \quad (3)$$

Rearranging the terms and subtracting a charge for the use of capital from both sides of the expression yields:

$$\text{Profits} \leq K \cdot (\gamma - W_1). \quad (4)$$

Allowed profits are a multiple of the rate-base capital. Since both the allowed rate of return (γ) and the price of capital (W_1) were assumed to be parameters for the firm, increasing allowed profits requires that the firm expand its rate base.

The utility will not be indifferent to the methods used to expand the rate base. Rate base expansion will proceed by allocating each additional unit of capital to its "most productive use," thereby increasing revenue or reducing noncapital costs. *The additional revenue or reduction in costs can in turn be used to purchase more capital inputs and expand allowed earnings even further.*

If the rate base were expanded by increasing output, still using the cost minimizing input combinations, the utility would move from using K^* units of capital to $\hat{K}_1$ units as shown in Figure 1.

However, as A-J demonstrated with the imposition of an effective constraint on earnings, the utility changes the input combinations making production more capital intensive than the cost minimizing input combination. The evasive actions reduce the maximum potential profits but shift the profit hill to the right, π_2 in Figure 2. Increased capital intensity at $\hat{K}_2$ increases the allowed earnings over those attainable by merely expanding output ($\hat{K}_1$).

Acquisition of completely nonproductive inputs such as gold doorknobs or payment of price premiums for equipment adds to the rate base without adding to revenue or reducing noncapital costs. Hence, they dissipate the "fund" of excess profits more rapidly than the acquisition of productive capital inputs. A result from either action is a new potential profit curve, π_3 lying inside the initial, cost-minimizing profit curve. The allowed earnings attainable when the utility purchases nonproductive capital inputs at $\hat{K}_3$ are less than those attainable by purchasing productive inputs at either $\hat{K}_1$ or $\hat{K}_2$.

A regulated firm subject to an effective constraint upon earnings will seek to dissipate excess profits by purchasing those rate-base

FIGURE 1

POTENTIAL PROFITS AND ALLOWED EARNINGS AS A FUNCTION OF THE RATE BASE (ASSUMING COST MINIMIZING INPUT COMBINATIONS)

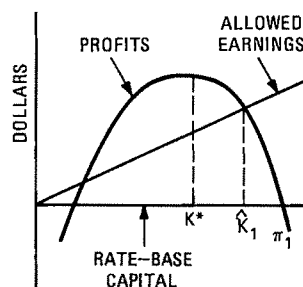
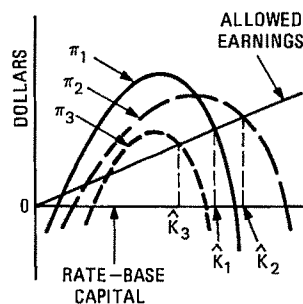


FIGURE 2

THE EFFECT ON THE POTENTIAL PROFIT CURVE OF ACQUIRING PRODUCTIVE AND NONPRODUCTIVE EXCESS CAPITAL INPUTS



⁵ The earnings constraint is effective if the constrained profits of the firm are less than the potential profits for an unconstrained firm in otherwise identical circumstances. Effective regulation does not necessarily imply an effective earnings constraint. Regulation may impose constraints upon the utilities with respect to the product, conditions of service, rate structure, and so forth in addition to the rate of return. Effective standards-of-service regulation by definition imposes additional constraints and costs upon the utility, lowers profits, and reduces the likelihood that an earnings constraint will be effective, *ceteris paribus*

additions which have the highest marginal revenue product. If the only opportunities open to the firm to dissipate excess profits were either to pad the rate base by purchasing capital goods with a zero marginal revenue product or to purchase capital goods with a negative marginal revenue product, the firm would choose the more profitable alternative and pad the rate base.⁶ Similarly, a utility subject to an actual-cost valuation regime will not voluntarily pay price premiums for equipment since the premiums have a zero marginal revenue product. Such alternatives would be attractive to a utility only if it did not have uses for capital offering a positive marginal revenue product.

The argument can be made with more rigor by considering the following mathematical model. An objective function for a regulated firm subject to rate-of-return regulation can be written:

$$\theta(K, V, \lambda, \mu) = R(K, V) - W_1 \cdot K \cdot \lambda - W_2 \cdot V - \mu[R(K, V) - W_2 \cdot V - \gamma \cdot B(K, \lambda) \leq 0]. \quad (5)$$

The variables in the objective function are restricted to the domains:

$$K > 0; \quad V > 0; \quad \lambda \geq 1; \quad 0 \leq \mu < 1.$$

Differentiating, the Kuhn-Tucker conditions for a maximum require that the variables satisfy the following equations.

$$[R_K^* - W_1 \cdot \lambda^* - \mu^* \cdot R_K^* + \mu^* \cdot \gamma \cdot B_K^*] \cdot K^* = 0 \quad (6)$$

$$[-W_1 \cdot K^* + \mu^* \cdot \gamma \cdot B_\lambda^*] \cdot [1 - \lambda^*] = 0 \quad (7)$$

$$[R_V^* - W_2 - \mu^* \cdot R_V^* + \mu^* \cdot W_2] \cdot V^* = 0 \quad (8)$$

$$-[R(K^*, V^*) - W_2 \cdot V^* - \gamma \cdot B(K^*, \lambda^*)] \cdot \mu^* = 0. \quad (9)$$

Equation (6) can be rearranged to show the A-J effect, except now the firm has two methods of dissipating profits—increasing the capital intensity of production and paying price premiums for capital goods:⁷

$$R_K^* = W_1 - \left[\frac{\mu^*}{1 - \mu^*} \right] \cdot (\gamma \cdot B_K^* - W_1) - \left[\frac{W_1 \cdot (1 - \lambda^*)}{(1 - \mu^*)} \right]. \quad (10)$$

Note that if the firm were to pay the competitive price such that λ^* were equal to one, the last term on the right hand side of the equation equals zero and (10) reduces to the usual A-J result. However, as presented, the equation indicates that some *combination of increased capital intensity and price premium* will maximize profits for the utility. The issue is what value for the premium factor will maximize profits and does it depend upon the method of rate base valuation?

The necessary conditions for a maximum will be met only when (7) is satisfied:

$$[-W_1 \cdot K^* + \mu^* \cdot \gamma \cdot B_\lambda^*] \cdot [1 - \lambda^*] = 0. \quad (7)$$

⁶ Zajac's contention that the firm will not pad its rate base [15], pp. 124-125, is only valid as long as the firm has alternative uses for capital inputs which offer positive marginal revenue products.

⁷ A profit maximizing utility will equate the marginal profitability of the last dollar expended on *each* type of evasive action. Where one type of evasive action by the regulated firm is found, one might reasonably expect to find other actions designed to satisfy the earnings constraint.

If the expression in the first set of brackets is not equal to zero, then $(1 - \lambda^*)$ must equal zero, implying that no price premium will be paid. Therefore, we are interested in the conditions under which the expression in the first set of brackets is equal to zero since when $[-W_1 \cdot K^* + \gamma \cdot \mu^* \cdot B_\lambda^*]$ equals zero the firm may be able to increase profits by paying a price premium.

With actual cost valuation the rate base can be rewritten as $K \cdot \lambda$ and $B_\lambda^* = K^*$. Rearranging the expression in the first set of brackets in (7) and setting it equal to zero yields:

$$K^* \cdot [\mu^* \cdot \gamma - W_1] = 0. \quad (11)$$

Since K^* has been assumed to be positive, (11) will equal zero only when $(\mu^* \cdot \gamma - W_1)$ equals zero. Solving (6) for μ^* and substituting into (11) yields:

$$K^* \cdot R_K^* \cdot (\gamma - W_1) = 0. \quad (12)$$

The allowed rate of return (γ) was assumed to exceed the cost of capital to the utility. Therefore, (12) will be satisfied only when the marginal revenue product of capital, R_K^* is equal to zero. For a monopolist, this implies that the marginal physical product of capital inputs must equal zero. Monopolists in general cannot be forced to expand output to the point of zero or negative marginal revenue and a regulated monopolist will prefer to expand the rate base as long as the marginal physical product of capital is nonnegative. A nonpositive marginal revenue product for capital implies a zero or negative marginal physical product for capital inputs; i.e., the firm must be operating in the Knightian Region III for capital inputs.⁸ *A necessary condition for a regulated monopoly subject to actual cost valuation to acquiesce to an equipment price premium is that its marginal physical product of capital must be nonpositive.*

The conditions under which a utility subject to reproduction cost valuation of the rate base will acquiesce to paying price premiums are naturally less stringent. Paying a higher price for one unit of new equipment may promote an upward revaluation of the entire stock of existing equipment.

Again, the results may be demonstrated mathematically by noting that the first bracketed term in (7) must equal zero, but the rate base has the form $B(K, \lambda)$. The substitutions now result in the following equation:

$$R_K^* \cdot (\gamma \cdot B_\lambda^* - W_1 \cdot K^*) + W_1 \cdot \gamma \cdot [K^* \cdot B_K^* - \lambda^* \cdot B_\lambda^*] = 0. \quad (13)$$

There are two different conditions satisfying (13) with a nonzero value for the marginal revenue product of capital:

$$\gamma \cdot B_\lambda^* - W_1 \cdot K^* = 0 \quad \text{and} \quad K^* \cdot B_K^* - \lambda^* \cdot B_\lambda^* = 0 \quad (i)$$

$$R_K^* \cdot (\gamma \cdot B_\lambda^* - W_1 \cdot K^*) \quad \text{and} \quad W_1 \cdot \gamma \cdot [K^* \cdot B_K^* - \lambda^* \cdot B_\lambda^*]$$

equal in absolute value but opposite in sign. (ii)

It may be possible for the utility to increase profits by paying price premiums even when the marginal revenue product of capital inputs are positive with reproduction cost valuation of rate-base capital.

Where regulation is combined with reproduction cost valuation of rate-base capital, the regulatory process may promote price con-

⁸ Westfield makes a similar contention [14], p. 419.

spiracies among equipment manufacturers. With actual cost valuation, the role of the regulatory process in fostering equipment conspiracies depends upon the plausibility of a utility operating in a Knightian Region III for capital inputs. The particular nature of the production process for fossil-fueled electrical generation developed in the following subsection indicates that operation in a Knightian Region III for capital is highly implausible.

□ **A process function for electrical power generation.** Approximately 80 percent of the electrical energy used in the United States was steam-electric generated in 1963.⁹ The generation process can be viewed as a series of transformations where the molecular energy of a fossil fuel is ultimately transformed into electrical energy. In the process only part of the fuel energy inputs will actually be transformed into electrical energy and the ratio of energy-out to energy-in is termed the “net thermal efficiency” of the transformation.

The designed level of thermal efficiency is an important property of the capital equipment used in the transformation process. Modern steam-electric generating units are designed to be operated only as integrated units, and the properties of the units may be considered rather than the properties of the individual components. The net thermal efficiency of the capital equipment (E) is treated here as an input for the process function.

Since energy cannot be created in the transformation process, there is some minimum energy input consistent with each level of energy output. The source of power in the production process is the fuel input (f , measured in Btu's per hour). The fuel input indicates the rate at which heat is uniformly supplied to the production process and instantaneously converted into electricity. The output of the process is the power produced by the generating unit (z , measured in kilowatts). There is a common physical constant for conversion between units:

$$3413 \text{ Btu's} = 1 \text{ kw hr.} \quad (14)$$

A simple process function for the relation between fuel inputs and kilowatts of output may be written:

$$z = \frac{f \cdot E}{3413 \text{ Btu's/kw hr}} \quad (15)$$

The expression simply states that if fuel is supplied at some uniform rate to a generating unit with a specific design thermal efficiency, a certain number of kilowatts will be produced.¹⁰

Equation (15) constitutes a process function for electrical power generation by a single generating unit. The capital inputs are implicitly represented in the net thermal efficiency of the transformation process and some unspecified capacity limitations. The properties of the capital inputs enter the production process in catalytic fashion. We assume that the firm does not attempt to operate the unit at a rate exceeding capacity and will not explicitly introduce the capacity

⁹ See [11], p. 26.

¹⁰ A corresponding cost function for a steam-electric generating unit may be found in Emery [3], p. 132.

property of the capital equipment into the process function. However, capacity limitations still apply to the generating unit.

An expression for the acquisition price of capital goods as a function of the properties specified in the process function is necessary to relate the process function to the economist's traditional dollar measures of capital inputs. Previous studies have found economies of capacity in purchasing large steam-electric units,¹¹ but the cost of increasing thermal efficiency increases as thermal efficiency increases.¹² Plausible relationships between capacity (X), thermal efficiency (E) and the acquisition prices of a steam-electric generating unit (K_G) are that:

$$\frac{\partial K_G}{\partial E} > 0; \quad \frac{\partial K_G}{\partial X} > 0; \quad \frac{\partial^2 K_G}{\partial E^2} > 0; \quad \frac{\partial^2 K_G}{\partial X^2} < 0. \quad (16)$$

One expression for acquisition price that will satisfy the stated conditions is

$$K_G = \alpha \cdot X^\beta \cdot E^\psi, \quad (17)$$

with $\alpha > 0$, $0 < \beta < 1$, and $\psi > 1$.

Increasing the thermal efficiency of the steam cycle requires higher steam temperatures, higher pressure, more reheat cycles, lower condenser temperatures, etc., all of which increase the cost of capital equipment. With present production techniques the net thermal efficiency will always be less than one regardless of the level of expenditures.¹³ With the specified process function and acquisition price function for generating equipment, a utility has a virtually unlimited source of investment alternatives which offer a positive marginal physical product for dollar units of capital, and a Knightian Region III will not be reached in steam-electric generation.¹⁴

To summarize, with original cost valuation of rate-base capital, a profit maximizing electrical utility has a negative incentive to join or acquiesce with a price conspiracy by equipment manufacturers. Only with replacement cost valuation might the regulatory process foster equipment price conspiracies. The regulatory process cannot be indiscriminately cited as encouraging conspiracies among suppliers of capital goods.

■ One implication of the model developed in Section 2 is that the incentive to seek out or negotiate the lowest possible prices for new capital equipment may differ between utilities subject to differing rate-base valuation regimes. Specifically utilities subject to actual cost valuation have a negative incentive to pay conspiratorially higher equipment prices, while utilities subject to reproduction cost methods might conceivably benefit from higher prices or not be injured.¹⁵ The difference in attitudes towards equipment prices between utilities

3. The effects of different rate-base valuation schemes on the acquisition prices of steam-electric generating units

¹¹ See [4] and [11].

¹² See [2].

¹³ See [2], pp. 87-140 for a discussion of the physics of steam cycles.

¹⁴ Note that the basic argument on the implausibility of achieving a non-positive marginal physical product of capital can be applied to the transportation process for electric power [8], pp. 24-30.

¹⁵ The degree of incentive for price premiums will be related to the likelihood of transforming the higher price on one piece of equipment into an upward re-

subject to differing valuation schemes might be manifest in at least two ways: a difference in the prices paid for similar equipment or a difference in the success of a suppliers' conspiracy to raise equipment prices within the two categories. The implications of the model with respect to a utility's *disposition* towards higher equipment prices is not dependent upon the existence of a suppliers' conspiracy, but *achievement* of price differentials may hinge upon such a conspiracy. Thus, for a time span generally acknowledged to include a period of effective collusion among the turbine-generator suppliers, two hypotheses were formulated:

- (1) Electrical utilities subject to reproduction cost regimes paid prices for capital equipment higher than the prices paid by utilities subject to actual cost valuation regimes for similar equipment.
- (2) The differential between prices paid for similar generating equipment during the periods of conspiracy and nonconspiracy among the equipment suppliers will be greater for utilities subject to reproduction cost valuation than for those subject to actual cost valuation.

The second hypothesis is less dependent on the *ceteris paribus* assumption implicit in the first hypothesis.

The procedure used to test the hypotheses was first to construct a regression model to explain the acquisition prices of a class of steam-electric generating units. Then if the type of rate-base valuation is an important factor in determining the cost of a unit as suggested in the previous section, the differences or residuals between the observed and estimated acquisition prices should systematically differ between units subject to reproduction cost valuation and those subject to actual cost valuation.¹⁶

□ **The regression model.** The regression model was based on the results of other studies¹⁷ and data made available to the author on a private basis by a major manufacturer of steam-electric equipment. As stated previously in (17), the acquisition price of a generating *unit* is primarily a function of two variables: manufacturer's rated output capacity (*CAPC*) and design thermal efficiency of the unit (*EFFC*).

$$K_G = \alpha \cdot (CAPC)^\beta \cdot (EFFC)^\psi. \quad (17a)$$

Problems with multicollinearity between the two variables forced independent estimation of the exponent on thermal efficiency, which was based on earlier work by the author.¹⁸ Values for ψ in the range $1 < \psi \leq 2$ imply that the marginal cost of thermal efficiency is positive and increasing but not increasing at an increasing rate. The limited range of thermal efficiencies found in the sample were not on the frontiers of technology at the time of the sample, and should not be subject to the very sharply increasing marginal costs found by

valuation of all rate base capital and the time lag involved before this can be achieved.

¹⁶ See Sloss [7], pp. 341–347.

¹⁷ In [2] and [4].

¹⁸ See Emery [3].

Cootner and Lof.¹⁹ The results are arbitrarily reported for $\psi = 1.2$, but they hold for values of ψ in the range of 1.05 to 2.0.²⁰

The generating units included in the final samples were coal-fired units, with a single reheat cycle, having a rated capacity of 100 megawatts or over and purchased by a utility.²¹ Units of this type corresponded to the units described in the manufacturer's private report. The Federal Power Commission series *Steam Electric Plant Construction Costs and Annual Production Expenses*²² for the period 1959–1967 was used as the primary source of cost and technical data. The contract dates on the turbine-generator units were obtained from an exhibit in the OVEC vs. GE case.²³

Generating units with the characteristics specified above and with turbine-generator contract dates between 1954 and 1961 were included in the initial sample. During 1955 and again in 1960 the so-called “sales” of equipment occurred as the “agreements” between suppliers were ineffective. Units with a contract date during the period of 1956–1959 were judged more likely to exhibit price differentials than those with contract dates during 1954–1955 and 1960–1961.

The name-plate capacity of the individual generating *units* reported by the FPC was used as the capacity observation. Since the FPC data did not provide direct information on the design thermal efficiency of the units, efficiency was estimated on the basis of the steam cycle characteristics (psi, temperature, reheat temperature). The observations on acquisition prices were the recorded equipment costs for the units exclusive of land, land rights, structures, and improvements. The difference in the recorded costs for year t_0 and t_{-1} was used as the cost of a unit added to an existing station.²⁴

Five other variables were included in the cost function: a variable related to the number of units in each plant (*UNUM*), area of the country in which the plant was located (*AREA*), year of turbine-generator contract (*YEAR*), a dummy variable if the unit used artificial sources of condenser cooling water (*CWTR*), and a dummy variable if the unit was of “outdoor” or “semi-outdoor” type of construction (*STRT*).

Values for the variable *UNUM* were calculated as one divided by the number of units in the plant. Since *UNUM* is the reciprocal of the number of units the regression coefficient should be positive indicating that there are economies of number in constructing steam electric generating stations.²⁵

AREA was assigned values between 1 and 3 based on the plant location within the major U. S. census regions. For plants located in the low construction cost area of the South, *AREA* was assigned a

¹⁹ In [2].

²⁰ Values of ψ greater than 2 noticeably reduce the coefficient of determination, increase the regression coefficient on capacity, and reduce the difference in the residuals between the different valuation regimes.

²¹ All units listing coal as the major fuel or as one of a number of fuels were classified as coal-fired plants.

²² See [12].

²³ In [5].

²⁴ This practice will understate the cost of an additional unit since normally some depreciation of the existing plant would have occurred during the year. The regression coefficient on *UNUM* will be biased downward.

²⁵ See Komiya [4].

value of 1; for the Midwest, 2; and the high construction cost areas of the Northeast and West, 3.

Positive values were assigned to the dummy variables *CWTR* and *STRT* whenever the units used artificial sources of cooling water or outdoor type construction respectively. The regression coefficients on *UNUM*, *AREA*, *CWTR*, and *STRT* are expected to have positive signs.

The regression equation can be written in logarithmic form as:

$$\log (COST/FFFC\psi) = \log b_0 + b_1 \cdot \log (CAPC) + b_2 \cdot \log (UNUM) + b_3 \cdot \log (AREA) + b_4 \cdot \log (YEAR) + b_5 \cdot \log (CWTR) + b_6 \cdot \log (STRT) + \log \mu. \quad (18)$$

Nine different regression equations were estimated for equation 18 using acquisition prices of generating units in different time periods as the dependent variables. The observations for each two-year period between 1954 and 1961 were grouped to enlarge sample sizes, and seven sets of coefficients were estimated. In addition, the coefficients for 1956–1959, a period of supposedly effective conspiracy and the coefficients for the entire time period, 1954–1961, were also estimated. The coefficients are presented in Table 2.

TABLE 2

REGRESSION COEFFICIENTS AND STANDARD ERRORS FOR STEAM-ELECTRIC GENERATING UNIT ACQUISITION COST FUNCTIONS 1954–61

	1954-55	1955-56	1956-57	1957-58	1958-59	1959-60	1960-61	1956-59	1954-61
CAPC (UNIT CAPACITY)	0.9116** (0.0708)	0.8638** (0.0620)	0.8620** (0.0696)	0.9109** (0.1087)	1.009** (0.0842)	0.9436** (0.0738)	0.8313** (0.0806)	0.9302** (0.0508)	0.9212** (0.0359)
UNUM (1/NUMBER OF UNITS)	0.0846 (0.0458)	0.1054** (0.0368)	0.0784* (0.0370)	0.1197 (0.0759)	0.1572* (0.0579)	0.1045 (0.0514)	0.1161 (0.0674)	0.1027** (0.0292)	0.1037** (0.0231)
AREA (PLANT LOCATION)	0.0873 (0.0528)	0.1140** (0.0422)	0.1678** (0.0523)	0.1081 (0.1279)	0.1953* (0.0909)	0.2120** (0.0654)	0.2586** (0.0904)	0.1592** (0.0446)	0.1489** (0.0313)
YEAR (INITIAL OPERATION)	0.0327 (0.0961)	0.1244* (0.0586)	0.0348 (0.0707)	-0.1958 (0.1473)	0.0944 (0.1430)	-0.2235* (0.0920)	0.0949 (0.1185)	-0.0308 (0.0223)	-0.0193 (0.0100)
CWTR (COOLING WATER SOURCE)	-0.0159 (0.0294)	-0.0308 (0.0229)	-0.0317 (0.0289)	-0.0303 (0.1349)	-0.0237 (0.0448)	-0.0147 (0.0341)	-0.0147 (0.0452)	-0.0187 (0.0236)	-0.0129 (0.0171)
STRT (OUTDOOR CONSTRUCTION)	-0.0102 (0.0288)	-0.0085 (0.0216)	0.0093 (0.0243)	-0.0085 (0.0855)	0.0467 (0.0427)	0.0161 (0.0286)	0.0082 (0.0336)	0.0198 (0.0206)	-0.0064 (0.0143)
COEFFICIENT OF DETERMINATION	0.82	0.80	0.86	0.87	0.92	0.92	0.91	0.88	0.86
F	35.89**	41.27**	31.69**	13.12**	29.88**	42.82**	24.38**	66.97**	132.33**
NO. OF OBS.	52	66	36	18	22	28	20	58	130

* SIGNIFICANT AT THE 0.05 LEVEL OR ABOVE ON A TWO-TAILED TEST
 ** SIGNIFICANT AT THE 0.01 LEVEL OR ABOVE ON A TWO-TAILED TEST.

The estimated coefficients for capacity (*CAPC*), number of units (*UNUM*), and geographical area (*AREA*) are consistently positive as predicted. The regression coefficient on capacity is very significant in each period, and with the exception of the 1956–1959 period, the estimates are in the neighborhood of those found by Komiya.²⁶ Subsequent analysis in Table 4 suggests that the discrepancies between the estimated coefficients on capacity for regressions number 4–6 and those for other time periods may be due to the suppliers' agree-

ments in effect during 1956–1959. The area coefficients are significant in most periods. The variable for the number of units in each plant yields significant coefficients for the periods 1956–1959 and 1954–1961, but the coefficients for the two-year periods are not consistently significant.

The estimated coefficients for structure (*STRT*) and cooling water treatment (*CWRT*) do not agree with *a priori* expectations. Neither variable consistently yields significant regression coefficients. The coefficient on cooling water consistently has a negative sign, contrary to predictions, which is puzzling. However, cooling towers, spray ponds, and wells are not “equipment” in the sense the word has been used above. While they all add to the total cost of the plant, they may reduce the acquisition costs of furnace-boilers and turbine-generators in some inexplicable manner.

There is some intercorrelation among three variables: number of units, structure, and cooling water treatment. Several of the zero-order correlation coefficients fall in the range 0.25 to 0.58. Multicollinearity may account for the switching in arithmetic sign on the regression coefficient for structure and the nonsignificance of the regression coefficient on *UNUM* in some instances.

The regression coefficient on the year of contract for the turbine-generator (*YEAR*) is significant in two, two-year periods, 1955–1956 and 1959–1960. The coefficient is positive in 1955–1956 indicating an increase in the price of equipment between 1955 and 1956, and negative in 1959–1960 indicating a reduction in equipment prices between those years. It is generally acknowledged that the equipment manufacturers’ agreements became effective again in 1956 following the “sales” of 1955. The first conspiracy indictments against the manufacturers were handed down in February of 1960.

□ **The residuals.** The residuals from regression equations (1) through (8) were grouped by regime and analyzed to determine whether the type of rate-base valuation influenced the acquisition prices of generating units. If the hypotheses stated earlier are correct, then the residuals or time pattern of residuals from units subject to some form of reproduction cost valuation should systematically differ from the residuals for units subject to actual cost valuation.

*Moody’s Public Utilities Manual*²⁷ was used as a basis for determining the geographical area served by each plant, and hence, the state with primary jurisdiction. Plants appearing to be subject to both types of valuation regimes due to joint ownership or the area served were excluded from the sample. States where the regulatory commission listed “reproduction cost,” “fair value,” or “all elements considered” as the rate-base valuation method were all classified in the reproduction cost category. The actual cost category included states where the valuation methods were listed as “original cost” or “prudent investment.” Plants located in states not regulating electricity or owned by municipalities or cooperatives or built by the TVA were included in the actual cost category.

During the time period included in the sample, there were numerous changes in the valuation methods listed by the state commissions with the Federal Power Commission.²⁸ Privately owned plants lo-

²⁷ [6].

²⁸ In [10].

TABLE 3

TEST STATISTICS ON THE DIFFERENCES IN RESIDUALS BETWEEN PLANTS SUBJECT TO ACTUAL COST AND REPRODUCTION COST VALUATION REGIMES 1954-61

A) EXCLUDING OBSERVATIONS FROM STATES LISTING A CHANGE IN VALUATION METHOD BETWEEN THE 1954, 1960 AND 1967 SURVEYS								
	1954-55	1955-56	1956-57	1957-58	1958-59	1959-60	1960-61	1956-59
1) "t"	-1.510	-1.772	0.5130	1.755	1.556	1.541	0.7633	1.811
(df)	50	64	34	16	20	26	18	56
2) χ^2	0.373	1.52	0.114	0.833	2.36	1.43	2.12	3.38
(df)	1	1	1	1	1	1	1	1
3) % OF RESIDUALS WITH "CORRECT" SIGN	46	42	47	56	64	61	40	62
B) USING 1967 CLASSIFICATION OF VALUATION METHODS AND INCLUDING OBSERVATIONS FROM STATES LISTING A CHANGE IN VALUATION METHOD*								
1) "t"	-1.1590	-1.5080	1.1257	2.2376**	1.4432	1.5047	0.7479	2.0625**
(df)	58	72	38	21	24	28	20	62
* NEW JERSEY CLASSIFIED AS AN ACTUAL COST STATE SINCE ORIGINAL COST IS TAKEN AS THE BEST MEASURE OF FAIR VALUE.								
** SIGNIFICANT AT THE 0.05 LEVEL OR ABOVE								

cated in any of the nine states listing one or more changes in valuation methods in the surveys for 1954, 1960, or 1967 were excluded from the sample in calculating the statistics shown in Part A of Table 3. The second set of *t*-statistics, Part B, includes observations from states reclassified as to valuation method.

If the first hypothesis—that utilities subject to reproduction cost valuation paid higher prices for similar units than did utilities subject to actual cost valuation—is correct, then the residuals from the units in the reproduction cost category should systematically exceed the residuals from units in the actual cost category. The test statistics, shown in Part A of Table 3, indicate no significant differences between the residuals of the two groups during any of the eight time periods included in the sample. The *t*-statistics are positive for all periods during the conspiracy, but not greater than zero at a 0.05 level of significance or above. Similarly, the *t*-statistic for the period 1956–1959 is positive but not significant. None of the χ^2 statistics are significant at the 0.05 level or above.

The statistical results are sensitive to the particular FPC classification of rate-base valuation methods used by state commissions. To the extent that court decisions and/or local options determine the valuation method, the stated classification by state commissions may lag the actual practices. The reclassification of one state, Missouri, into the "reproduction cost" category in the 1967 FPC report²⁹ increases the values of the *t*-statistics to the 0.05 level of significance for the time periods 1957–1958 and 1956–1959 as shown in Part B of Table 3. There were no appreciable changes in the χ^2 statistics due to reclassification.

New Jersey was officially reclassified from the original cost category to "fair value" in the 1967 FPC report, but the classification was qualified such that original cost was taken as the best measure of fair value. Therefore, New Jersey was left in the actual cost

²⁹ *Ibid.*

category even though the decision reduced the values and significance of the *t*-statistics for the conspiracy period. The residuals from units in certain states were consistently large in value and opposite in arithmetic sign to that predicted. Deviant residuals were the rule in Ohio and North Carolina in the reproduction cost category, New York in the actual cost category, and Missouri and New Jersey among states experiencing a change in classification.

The *t*-statistics are noticeably higher during the 1956–1959 period than they are for the other years. Similarly, the percentage of residuals with the predicted or “correct” arithmetic sign improves noticeably during the 1956–1959 period. However, only under the 1967 classification of valuation methods do the *t*-statistics allow rejection of the null hypothesis and then only for certain periods.

If the second hypothesis is correct, the increase in plant prices due to conspiracy should be greater for utilities subject to reproduction cost valuation than for those subject to actual cost valuation. The residuals from regression (9) were cross-classified into actual and reproduction cost categories and by 1956–1959, versus other years in the period 1954–1961. The residuals were then analyzed using analysis of variance techniques.³⁰

TABLE 4
ANALYSIS OF VARIANCE ON RESIDUALS FOR REGRESSION EQUATION 9, 1954–61
CROSS CLASSIFIED BY VALUATION METHOD AND TIME PERIOD

SOURCE OF VARIATION	SUM OF SQS.	<i>df</i>	MEAN SQUARE	<i>F</i> _{1,125}
VALUATION METHOD	0.0000264	1	0 0000264	0.04
TIME PERIODS	0.01033	1	0.01033	16.72*
INTERACTION	0.00167	1	0.00167	2.704
ERROR			0.000617	
* SIGNIFICANT AT THE 0.01 LEVEL OR ABOVE.				

The results shown in Table 4 indicate that there were very significant differences between the means for the two time periods. Prices paid by utilities subject to actual cost valuation were approximately 6 percent higher during the 1956–1959 period than during the other four years of the period, but the difference is not significant at the 0.05 level. Prices paid by utilities subject to reproduction cost valuation were 15 percent higher during the 1956–1959 period than during the other four years, and the difference is significant at the 0.01 level. On the average, prices increased 10 percent during 1956–1959 when compared to the two, two-year periods immediately preceding and following a period of supposedly effective conspiracy.

Neither the valuation method nor the interaction between valuation method and the time periods accounted for a significant amount of variation. The low *F*-statistic on interaction implies that the changes between the means for residuals due to the time classification does not differ significantly between regimes. Therefore, we cannot conclude that the increase in plant prices due to conspiracy was greater for utilities subject to reproduction cost valuation than for those subject to actual cost valuation. While the price increase was very

³⁰ See [13], pp. 381–382.

significant for utilities subject to reproduction cost valuation and not significant for utilities subject to actual cost valuation, the *difference* in the increase is not significant.

The significant increase in prices to utilities subject to reproduction cost valuation does not show up as higher prices for generating units in this group because utilities subject to reproduction cost valuation paid lower prices for plants both before and after the conspiracy period than did those utilities subject to actual cost valuation. This fact tends to cloud the interpretation of the results for the initial hypothesis.

The timing of contracts on turbine-generator units does not suggest that the utilities encouraged or acquiesced to the price increases. Contracts for one-third of the plants in the sample were awarded in 1955, a "sales" year and 45 percent of all observations came from 1955 and the first half of 1956. Utilities subject to reproduction cost valuation did not account for a disproportionate share of the units ordered during the 1956–1959 time period which might be expected if they were encouraging or acquiescing to the conspiracy. The pattern of orders suggests that all utilities sought to avoid the higher prices in effect during the early 1950s and again in 1956–1959 by waiting for breaks in the agreements among suppliers.

4. Conclusions

■ When an effective earnings constraint is combined with any form of reproduction cost valuation scheme the regulatory process can legitimately be censured for encouraging utilities to join or acquiesce with equipment price conspiracies. But given the nature of the production process for steam-electric generation, utilities subject to actual cost valuation schemes have a negative incentive to join such conspiracies. The regulatory process *per se* does not encourage price conspiracies among equipment suppliers. The fault if any lies with the regulatory commissions, statutes, or courts that specify the use of reproduction cost or "fair value" rate-base valuation schemes.

The empirical evidence, though mixed in nature, does not allow acceptance of a hypothesis that utilities subject to some form of reproduction cost valuation paid higher acquisition prices for similar plants than did utilities subject to some form of actual cost valuation. Only with the 1967 FPC classification of valuation methods by states do significantly higher prices appear on generating units purchased by utilities subject to reproduction cost valuation.

Prices were significantly higher during 1956–1959 than they were for the two, two-year periods immediately preceding and following the period. The increase in prices for units purchased by utilities subject to reproduction cost valuation, 15 percent, was very significant; but the increase for units purchased by utilities subject to actual cost valuation, 6 percent, was not significant. However, again the null hypothesis of no difference in the increases between units purchased by utilities in differing valuation regimes cannot be rejected.

Finally, it should be noted that reproduction cost valuation expands the number of evasive actions that may be taken by a utility to combat an effective constraint upon earnings. The utility can inflate the value of the rate base by paying conspiratorially higher prices for new capital equipment or it may seek to influence subjective rate base valuation through legal or political action. It is doubtful

that these types of evasive actions promote economic efficiency or social welfare to a greater extent than do the reduced number of options available under actual cost valuation.

References

1. AVERCH, H. AND JOHNSON, L. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol. 52, No. 5 (December 1962), pp. 1053-1069.
2. COOTNER, P. AND LOF, G. *Water Demand for Steam-Electric Generation*, Baltimore: Johns Hopkins Press, 1965.
3. EMERY, E. D. "An Investigation of the Potential Welfare Effects Associated with Rate-of-Return Regulation of the Steam-Electric Industry in the United States." Unpublished Ph.D. dissertation, University of Minnesota, 1969.
4. KOMIYA, R. "Technological Progress and the Production Function in the United States Steam Power Industry." *Review of Economics and Statistics*, Vol. 44, No. 2 (May 1962), pp. 156-166.
5. OVEC vs. GE (Exhibit). "Turbine Generator Transactions: 1948 through 1962."
6. ST. CLAIR, F. J., ed. *Moody's Public Utilities Manual: 1968*. New York, 1968.
7. SLOSS, J. "Regulation of Motor Freight Transportation: A Quantative Evaluation of Policy." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 327-366.
8. SMITH, V. *Investment and Production: A Study in the Theory of the Capital Using Enterprise*. Cambridge: Harvard Univ. Press, 1966.
9. TAKAYAMA, A. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol. 59, No. 3 (June 1969), pp. 255-260.
10. U. S. FEDERAL POWER COMMISSION. *Federal and State Commission Jurisdiction and Regulation: Electric, Gas and Telephone Utilities*. Washington: U. S. Government Printing Office, 1954, 1960, and 1967.
11. ———. *National Power Survey: Volume 1*. Washington, D. C.: U. S. Government Printing Office, 1964.
12. ———. *Steam-Electric Plant Construction Costs and Annual Production Expenses*. Washington, D. C., 1959-1966.
13. WALKER, H. M. AND LEV, J. *Statistical Inference*. New York: Holt, Rinehart and Winston, 1953.
14. WESTFIELD, F. "Regulation and Conspiracy." *The American Economic Review*, Vol. 55, No. 3 (June 1965), pp. 424-443.
15. ZAJAC, E. "A Geometric Treatment of Averch-Johnson's Behavior of the Firm Model." *The American Economic Review*, Vol. 60, No. 1 (March 1970), pp. 117-125.

Depreciation in a simple regulatory model

Bruce L. Jaffee

Assistant Professor of Business Economics and Public Policy
Indiana University

This study introduces depreciation as an explicit variable in a regulatory model. The goal of the firm is assumed to be maximization of discounted cash flow. For simplicity it is assumed that the firm is required to use the same type of depreciation for book and tax purposes. It is shown that under the assumptions of the model, rate-base regulation strongly affects the type of depreciation which the firm would consider optimal. More specifically, the solution of the model indicates that in order to maximize discounted cash flow while satisfying the regulatory control, the firm should allocate depreciation in each period so that it will be an increasing function of time. As a result, undiscounted cash flow will increase throughout the time horizon.

1. Introduction

■ With few exceptions, most models of the regulated firm ignore all aspects of depreciation. Averch and Johnson, for example, state that “for simplicity, we assume that depreciation . . . is zero. . . .”¹ The main purpose of this study is to develop a simple regulatory model which explicitly includes depreciation.

In order to show clearly the effects of depreciation, we suppress the effect of regulation on the level of output and on the capital-labor ratio, issues which have received considerable attention in the literature.² We concentrate exclusively on the question of the optimal allocation of depreciation in a regulatory model. Furthermore, we ignore the problem of capital replacement. Instead we consider only the purchase of a single “lump” of capital equipment. Important contributions in the capital replacement field have been made by Alan Manne and Philip Tasch.³

Peter Linhart has made a notable contribution on the role of depreciation in a regulated context.⁴ His approach is quite different from the one developed here, a major difference being that the model discussed below has an explicit regulatory control. Linhart subsumes the control into a “revenue requirements” equation. Although they do not develop formal models, several authors have discussed the type of depreciation that firms should use—or be forced to use. Ben Lewis maintains that regulated firms should be required to use ac-

I wish to thank W. H. Oakland and especially F. T. Sparrow for helpful assistance with this research. I am also grateful to William F. Sharpe and an anonymous referee for comments on an earlier draft.

¹ In [1], p. 1054.

² For an excellent survey see Baumol and Klevorick [2].

³ In [7] and [8], respectively.

⁴ In [6].

celerated depreciation since it lowers their tax liabilities, and hence the costs on which rates are set.⁵ Sidney Davidson discusses the question of whether or not firms which use accelerated depreciation should establish a fund for the higher taxes which, it is alleged, will accrue in the future if accelerated depreciation is instituted. He maintains that except in unusual cases the taxes of a firm will not be increased in the future if accelerated depreciation is substituted for straight-line depreciation.⁶

■ In the model presented below, time is discrete, taxes are explicitly included, the same type of depreciation is required to be used for book and tax purposes, and depreciation is a deductible expense for tax purposes. Following most of the financial literature, we assume that the objective function of the firm is maximization of discounted cash flow. The goal of the firm is to find the time path of depreciation which maximizes discounted cash flow during the time horizon under consideration which extends from $t = 0$ to $t = T$. We assume that no institutional considerations limit the choice of depreciation paths except for certain feasibility constraints. In reality only a finite number of paths may be permitted by laws or accounting conventions. During each period in the time horizon the firm receives a constant amount of revenue from the fixed amount of capital equipment, x_1 , the single factor of production. We symbolize this revenue by $R(x_1)$.⁷ The actual level of x_1 is assumed to be determined outside of this model. Three costs are incurred in each time period. These are the annual cost of financing the capital equipment, crx_1 , (c is the purchase price of the capital equipment and r is the interest rate), depreciation, d_t , and income taxes. The second cost element is a non-cash expenditure and thus does not appear in the objective function. The complete model is

$$\text{Max}_{d_t} \left[D.C.F. = \sum_{t=0}^T \left(\frac{R(x_1) - crx_1}{(1+r)^t} \right) - T_o \sum_{t=0}^T \left(\frac{R(x_1) - crx_1 - d_t}{(1+r)^t} \right) \right], \quad (1)$$

subject to:

$$R(x_1) - d_t - T_o[R(x_1) - crx_1 - d_t] \leq s(cx_1 - \sum_{\tau=0}^t d_\tau), \quad t = 0, 1, \dots, T, \quad (2)$$

$$\sum_{t=0}^T d_t = cx_1, \quad \text{and} \quad (3)$$

$$d_t \geq 0. \quad (4)$$

The goal is to find the depreciation path, a discrete path symbolized

⁵ In [5].

⁶ In [3].

⁷ Alternatively we can consider the constant $R(x_1)$ to stand for total revenue net of all noncapital costs and to include economic depreciation (i.e. deterioration of the productive capacity of the capital equipment over time). The key assumption is that this revenue, net revenue, or net revenue with economic depreciation, symbolized by $R(x_1)$, be constant for all periods in the time horizon.

2. The model

by d_t , which maximizes discounted cash flow (*D.C.F.*), equation (1), while satisfying the feasibility constraints, equations (2)–(4). Note that in this formulation the optimal depreciation path is also the one which minimizes the present value of income taxes (subject, of course, to the constraints). In this model cash flow equals revenue net of capital costs (cx_1) and income taxes. We assume that these taxes in each period are a constant proportion, T_o , of revenue net of capital costs and depreciation. We assume that the discount rate and the annual percentage cost of financing capital are identical. Allowing them to differ would not materially affect the results. In the objective function, revenue, capital costs, and depreciation are all assumed to begin in period 0 and end in period T . Other assumptions, such as allowing depreciation or capital costs to start at $t = 1$, are possible but would not materially affect the results discussed below. The regulatory control, equation (2), states that revenue net of depreciation and taxes may not exceed s percent of the undepreciated rate base. Equation (3) states that the total of the depreciation taken in the time horizon exactly equals the cost of the equipment,⁸ and equation (4) requires that depreciation be nonnegative in each time period.

3. The solution

■ In the interest of brevity only the key steps and final results for the solution of the model represented by equations (1)–(4) are presented below.⁹ Note first that the objective function consists of two summations. The first is a constant; the second is maximized when

$$d_t = \begin{cases} cx_1 & \text{if } t = 0 \\ 0 & \text{if } t \neq 0 \end{cases},$$

but this solution is infeasible since it violates the regulatory control, equation (2). In general, the objective function encourages high depreciation during the early periods because depreciation appears in that function with a positive coefficient that decreases over time.

Turning to the regulatory control, we note that it can be rewritten as

$$(1 - T_o)R(x_1) + T_o cx_1 - (1 - T_o)d_t \leq s(cx_1 - \sum_{\tau=0}^t d_\tau);$$

$$t = 0, 1, \dots, T. \quad (5)$$

An increase in d_t , say Δd_t , lowers both sides of the inequality, with the left-hand side changing by $-(1 - T_o)(\Delta d_t)$ and the right-hand side changing by $-s(\Delta d_t)$. Since T_o is the marginal (federal and state) income tax rate (usually between 0.45 and 0.60) and s is the maximum allowed rate of return on capital (usually between 0.05 and 0.10), the left-hand side of equation (5) will fall considerably faster than the right-hand side when d_t is increased (between 4.0 and 11.0 times faster, based on these figures). Nonetheless, there must be enough depreciation remaining (from the total cx_1) in each time

⁸ Of course, the possibility exists that the accounting life of the capital equipment, call it K , differs from the length of the planning horizon. However, for simplicity we will assume that $K = T$.

⁹ The detailed solution is available from the author. Also see Jaffee [4], especially chapter 4.

period so that equation (5) can hold. Otherwise the solution will not be feasible. Using equation (5) recursively, the following existence condition for the model can be developed:

$$-\frac{K}{s} \left[1 - \left(\frac{1 - T_o}{1 - T_o - s} \right)^{T+1} \right] \leq cx_1, \quad (6)$$

where $K = (1 - T_o)R(x_1) + T_o crx_1 - scx_1$.

This equation is the existence condition. If it is satisfied, there will exist a feasible solution to the model described by equations (1)–(4). Note that the term in brackets approaches $-\infty$ as $T \rightarrow +\infty$. In such a case equation (6) can never be satisfied if $K > 0$: there will exist no solution to the problem.

Assume that equation (6) is satisfied so that there is a feasible solution to the problem. Then using equations (3) and (5) and the fact that the objective function, equation (1), encourages rapid depreciation, the following optimal feasible solution to the model can be derived easily:

$$d_t^* = \begin{cases} \hat{d}_t = R(x_1) + \frac{T_o crx_1}{1 - T_o} & (\text{for } t = T) \\ \hat{d}_t = \left(1 - \frac{s}{1 - T_o} \right) d_{t+1}^* & (\text{for } t = 1, 2, \dots, T-1) \\ cx_1 - \sum_{t=1}^T d_t^* & (\text{for } t = 0). \end{cases} \quad (7)$$

Here $\hat{d}_t$, $t = 0, 1, \dots, T$ represents the minimum amount of depreciation that must be taken in period t in order to satisfy equation (5), and d_t^* , $t = 0, 1, \dots, T$ represents the optimal level of depreciation for that time period. Since $0 < \left(1 - \frac{s}{1 - T_o} \right) < 1$, d_t ($t = 1, 2, \dots, T$) is an *increasing* function of t . Furthermore, since d_0^* must equal or exceed $\hat{d}_0$, which in turn is less than d_1^* , we must have $\hat{d}_0 \leq d_0^* \leq d_1^*$. Figure 1 sketches two optimal feasible depreciation paths: the first illustrates the case where $d_0^* = \hat{d}_0$; the second where $d_0^* > \hat{d}_0$.

■ The solution of this model is somewhat surprising. From the point of view of the rate-base regulated firm whose objective is maximization of discounted cash flow, the optimal depreciation path is one in which—with the possible exception of the initial time period—more depreciation is taken in the later periods. Regulation may thus induce the firm to use the reverse of accelerated depreciation. In terms of cash flow, the solution shows that undiscounted cash flow increases through the time horizon. Whether discounted cash flow increases or decreases depends upon the magnitude of the discount rate.

References

1. AVERCH, H. AND JOHNSON, L. L. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol. 52, No. 5 (December 1962), pp. 1053–1069.

4. Conclusion

2. BAUMOL, W. J. AND KLEVORICK, A. K. "Input Choices and Rate of Return Regulation: An Overview of the Discussion." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 162-190.
3. DAVIDSON, S. "Accelerated Depreciation and the Allocation of Income Taxes." *Accounting Review*, Vol. 33, No. 2 (April 1958), pp. 173-180.
4. JAFFEE, B. L. "Aspects of the Regulated Public Utility: Misallocation, Marginal Cost Pricing, and Depreciation." Unpublished Ph.D. thesis, Johns Hopkins University, Baltimore, 1971.
5. LEWIS, B. "The Duty of a Public Utility Is to Reduce Its Income Tax Liability by Using Accelerated Depreciation." *Land Economics*, Vol. 35, No. 2 (May 1959), pp. 104-114.
6. LINHART, P. B. "Some Analytical Results on Tax Depreciation." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 1 (Spring 1970), pp. 82-112.
7. MANNE, A. S. "Capacity Expansion and Probabilistic Growth." *Econometrica*, Vol. 29, No. 4 (October 1961), pp. 632-649.
8. TASCH, P. "Optimal Capital Policies for a Regulated Firm." Unpublished Ph.D. thesis, Johns Hopkins University, Baltimore, 1969.

Television's second chance: a retrospective look at the Sloan Cable Commission

Harvey J. Levin

Augustus B. Weller Professor of Economics
Hofstra University
and
Senior Research Associate
Center for Policy Research

This article examines the impact of an independent, nongovernmental commission's study on the regulatory policy governing cable television. In 1971, the Sloan Commission on Cable Communications released a final report—the Sloan Report—after privately assessing cable television's growth potential, cable's means to cross subsidize public channels, and cable's organization in terms of cross ownership and common carrier status. In addition to the brief Report, the Commission produced background papers and internal memoranda, which underscore the study's value in clarifying issues and in helping to structure the field for the future in the face of "regulatory hangover."

■ "Regulatory hangover" is the euphemism which refers to the failure of the regulatory framework to respond to technological change. The author will examine this phenomenon in the context of cable television, particularly with respect to the deliberations of a nongovernmental group, the Sloan Commission on Cable Communications.¹ As an independent citizens' group with eminent qualifications and full access to outside expertise, the Sloan Commission should have been better able than the Federal Communications Commission to discard the constraints of outmoded doctrine, and to devise rules more in the tune with the new technology. Was this in fact the case?

In seeking a balanced answer to this question, the reason for a nongovernmental commission must be kept in mind. First, such a private entity affords an independent delineation and audit of key

1. Introduction

This paper was researched and drafted with help from the National Science Foundation which the writer gratefully acknowledges. Special thanks must go to colleagues at Stanford, Yale, Berkeley, Harvard, and Chicago; and to staff members of the Office of Telecommunications Policy and the FCC, and to the Sloan Commission for access to many of the Sloan Commission's unpublished papers without which any fair evaluation would have been far more difficult. However, none of the conclusions reached herein, for which the author assumes full responsibility, is necessarily shared by any of these individuals or organizations.

¹ In addition to a Final Report, *On the Cable: The Television of Abundance* (1971) (henceforth called *Sloan Report*), Sloan commissioned some twenty background papers that were eventually released, and others that were not. There were also twelve private meetings of the Commission with invited witnesses (reported in unpublished Minutes), and a number of major internal memoranda.

issues so often riddled with conflicting interests. Second, proposals too sensitive for formal governmental airing may be screened for political soundness and viability and then passed on to high ranking public officials.

Notwithstanding the frequent criticism that the Sloan Commission bungled a unique opportunity to clarify issues and to help structure the field for the future, this author holds that, despite the *Final Report's* brevity and journalistic style, both it and surely the background papers contain valuable information for newcomers to the cable controversy. To do justice to the inquiry, the author has gone beyond the *Final Report* to consider some of the more relevant of these background papers, although his critique will necessarily focus on omissions and deficiencies of the *Final Report*.

Perhaps the major Sloan contribution was to bring special attention to the FCC ban on distant signal importation by cable into leading broadcast markets; the economic consequences of relaxing that ban; and the case for mandatory cross subsidy when cable penetration is sufficiently high. The vague, elusive character of FCC program service standards has long made a mockery of efforts to negotiate back from recalcitrant licensees, public service contributions commensurate to the value of the license privileges conferred. Whatever else one may say about economic welfare implications, Sloan unquestionably gave status to at least one thought-provoking way out of this dilemma—the mandating of special purpose channels as an alternative to current broadcast regulation “by raised eyebrow.”

In examining Sloan's role I shall therefore focus on three links—its scrutiny of cable's growth potential, cable's wherewithal to cross subsidize public channels, and organizational structure with special reference to cross ownership and a common carrier approach.

2. Growth of cable

■ The Sloan Commission saw cable as a virtual panacea for roadblocks of the spectrum era. Equating channel capacity with cable's unique potential for social good,² Sloan examined numerous barriers to cable growth. Yet the investigation for some reason failed to grasp the full relevance to that growth of our earlier disappointing experience with UHF broadcasting. It also failed to probe in sufficient depth the broadcast priorities to safeguard which FCC has overtly braked cable growth, and in particular the bearing of released spectrum as a possible factor in such growth. On each count the analysis seems needlessly impaired.

□ **Cable's abundance and the UHF analogy.** In broadcasting, the *potential* per market maximum of TV outlets ranges from 10 to 15 channels, whereas in 1970 the *actual* per market average of authorized TV stations in 49 leading markets was barely five,—with at least one UHF channel. In contrast, current cable systems are mainly 12-channel, but a leap to 20 channels is economically attractive and 40- or even 80-channel systems are now conceivable.

Will cable's potential copiousness materialize? As in the UHF case, subscribers will clearly not subscribe before attractive programming is generated, yet without significant cable penetration and subscriber income, how can such programming be financed? Sloan

² *Sloan Report*, pp. 43-46.

found an immediate answer in distant signal importation. To spread the program mix of New York City and Los Angeles more widely would be an important first step, though new conceptions were also needed. The Commission actually predicted an "achievable and conservative" cable penetration covering 40-60 percent of all homes by 1980 with 20 to 40 channels or more, limited two-way capacity, and extensive interconnection.³ The precise penetration ultimately attained would depend on just when interim penetration would become high enough to facilitate viable local program originations. At that moment subscriptions should soar all the more with low-cost satellite interconnection, and the chicken/egg syndrome might then be broken.⁴

Cable's contribution to diversity and merit service would depend in large measure on the extent and speed of this growth. The potential economic surplus which could cross subsidize public access, educational, and governmental channels and even some local originations, would reflect at least two things: the alleged monopoly character of local cable systems, and effective cable penetration.⁵

How swiftly will cable grow from nine percent penetration at present to the 20-percent level at which observers expect local originations to become viable,⁶ and to still higher levels nationally? The UHF episode offers some neglected guideposts to what must be at best a tentative answer. There too was a much heralded technological breakthrough whose promise failed to materialize. UHF stations left the air in droves in the early 1950s, partially because of their relatively deprived access to the leading networks' programming, the inadequacy of nonnetwork syndications as a substitute, UHF/VHF signal disparities, and the absence of click tuning. As a consequence, there was less incentive for viewers to convert their sets, for advertisers to spend money on UHF coverage, for networks to affiliate with UHF stations, *ad infinitum*. UHF therefore never really "got off the ground" and is still laggard today.⁷

In cable the question is whether copyrighted broadcast programs can be carried without inequity or competitive harm to broadcasting to facilitate cable growth and eventual program origination.⁸

What lessons could Sloan have drawn from the UHF experience? Faster more decisive regulatory action might have helped activate many of those still unused UHF frequencies in a variety of ways.

³ *Ibid.*, pp. 36-42, 173-74.

⁴ *Ibid.*, p. 38.

⁵ See generally, *Prospects and Policies for CATV*, Sept. 1971 (Sloan Paper by J. McGowan, R. Noll, and M. Peck), especially pp. 2-16, 49-63; *Common Carrier Access to Cable Communications*, March 1971 (Sloan Paper by L. Kestenbaum), pp. 5-8; Comments of U. S. Dept. of Justice, Docket No. 18397, Sept. 5, 1969, pp. 5-11; Comments of U. S. Dept. of Justice, Docket No. 18397-A, Dec. 7, 1970, pp. 17-20. See further discussion below, footnotes 14-15, 24, 25, 27-30, 41, and associated text.

⁶ Distant signal importation is widely viewed as a catalyst. If cable must pay for programming which is directly competitive with local broadcast service, and for independent but not network programs, signals or noncompetitive local station carriage, then the imported signals could be promotive of cable growth with no unfair competition with broadcasting. The wherewithal and incentives for cable interconnection and local origination would follow. See *Sloan Report*, pp. 38-39, 51-58; also *Prospects and Policies for CATV*, pp. 7-8.

⁷ See [6], pp. 341-47.

⁸ See in particular, *Report on Allocation of TV Channels to Senate Commerce Committee*, 85th Cong. 2d Sess. (1958), pp. 95-104. Also see [6], pp. 347-50.

Had we had the fortitude to relocate twenty pioneer VHF outlets, reassigning them to the UHF band, we could then have switched the whole VHF service to UHF. We could have further adopted the allocation plans proposed by Dumont and ABC to afford those networks access to VHF stations in the top 100 markets and to make a number of homogeneous UHF-only markets. The impoverished UHF stations could have been guaranteed compulsory access to NBC or CBS programs. The production of all-channel receivers could have been made mandatory much sooner. More decisive action could also have been taken on Bell's pricing as a factor in Dumont's withdrawal from TV networking. Or reduced-rate interconnection could have been mandated to facilitate the entry of a new UHF network. Was Sloan cognizant of analogous problems and opportunities in cable today?

Clearly Sloan *was* aware of the danger of discriminatory vertical access to monopolistic cable systems as between different programmers.⁹ But careful analysis of Dumont's restricted access to choice VHF stations in the top VHF markets, and of the impoverished UHF broadcaster's inability to secure popular CBS/NBC programs, might have better articulated the delicate chicken/egg character of cable growth. Until recently, FCC limitations on distant signal importation imposed constraints (comparable to those in the UHF episode) both on the revenues the potential broadcast exporter needed, and on the subscriber penetration for the cable importer. As in the UHF case, deficient software (programming) has slowed down the growth of hardware, whereas the lag in hardware and subscribers necessarily limits the funds available for exciting new programming.

The ban on signal importation has now been relaxed and there is no handicap anywhere in cable truly comparable to UHF's technical inferiority, the absence of click tuning, and the VHF/UHF intermixture problem. Yet the FCC's recent decision to extend exclusivity to nonnetwork broadcast programs in the top 50-100 markets¹⁰ could make it harder than otherwise to secure enough attractive programs to get cable "off the ground." A more explicit analysis of the earlier UHF episode might on these and other counts have clarified the current impasse. It could in any case have illuminated a better approach to two key questions: (1) how far will imported broadcast signals carry us beyond the current nine percent penetration level; and (2) will the new penetration level in fact suffice, along with pay-TV, to make local program originations or interconnected cable networks profitable, and thus to generate a truly national alternative to TV broadcast networks?

□ **Broadcast priorities and cable growth.** From the very outset the FCC has tried to control the rate and incidence of cable growth. The goal has been to phase in cable with a minimal impact on broadcasting. Critics see this as an unseemly reflection of the broadcaster's political power. But FCC's protective stance towards broadcasting is explicitly grounded on the public service responsibilities of broadcast licensees, and the question is how valid the Commission's reasoning is.

⁹ See above citations in footnote 5.

¹⁰ FCC Cable Rules, Fed. Reg. Vol. 37, No. 30, Pt. IV, Feb. 12, 1972, paras. 97-106.

Congress and the FCC have traditionally tempered regulatory reforms to sustain broadcasting's alleged wherewithal for public service,¹¹ doing so most dramatically in the network rules of 1941, again in 1958, still more recently in an attempt to reduce group ownership in the leading markets, and in cross ownership more generally. The prime-time access rule and TV network program procurement have also been handled in like fashion.

Likewise speedy cable progress has been braked deliberately to safeguard UHF broadcasting, local live service, and a wide geographic dispersion of broadcast facilities¹² and further to safeguard VHF monopoly rents and the news and public affairs programs they make it possible to subsidize. The FCC's sensitivity to broadcast equities is apparent, finally, in the back-tracking of its so-called consensus agreement.¹³ In contrast, public equities are implicit in the mandating of special purpose channels as a *quid pro quo* for distant signal importation. But how consistent was the Commission's economic logic? This is the context in which the Sloan Report must be evaluated.

To phase in distant signal importation consistent with broadcast structure and performance, the FCC has now permitted entry into the *second* fifty markets. Yet the major VHF rents are incurred in the *top* fifty, and it is there also that cable penetration could generate the largest surplus (given cable's local monopoly character). However, the skeptic cannot have it both ways: either the VHF rents exist and cable threatens them (in which case penetration ratios will be high and conducive to a large surplus), or cable does not seriously threaten the rents because the broadcast service is so distinctive (in which case the penetration ratios and the surplus will be much smaller).¹⁴

In the latter case, the FCC's protective stance is gratuitous, but even in the former case there is another compelling argument to let cable in. If one looks at *television as a whole*, both off-the-air and on-the-cable, the question is whether the reduction of VHF rents (and of off-air public service), is offset by (1) an increase in the cable surplus; (2) the structuring of greater cable diversity via mandated

¹¹ A seminal treatment of the basic issues appears in FCC, *Public Service Responsibilities of Broadcast Licensees* (1946). See also *Sloan Report* (1971), pp. 73-74 (effects on quality).

¹² See *Innovation and the Regulatory Agency: FCC's Reaction to CATV* (prepared for Sloan Commission by H. Goldin, Aug. 1970), pp. 5-26.

¹³ The "consensus agreement" appears in FCC Cable Rules, paras. 61-67. For a penetrating critique see Statement of Commissioner Nicholas Johnson Concurring in Part and Dissenting in Part in Dockets No. 18397, 18397-A, 18373, 18416, 18892, 18894.

¹⁴ These propositions must be qualified, however, where much cable income comes from subscriber TV (STV) channels. In that case cable might do well economically without seriously hurting broadcast advertising or VHF monopoly rents. Thus in February 1972, one commercial minute on prime time network TV was worth about \$2 per 1000 viewers to an advertiser, or \$12 for six commercial minutes per hour, or 1.2¢ an hour per viewer. But viewers paid between \$1 and \$2 per hour program in Hartford's pay-TV experiment in 1962-64. So that viewers individually were willing to pay direct some 100 to 200 times more than advertisers now pay the networks to reach them. Therefore, even a small diversion of TV network viewers to a cable/STV channel could generate far more for cable than it takes from broadcasting. Nevertheless, the above qualification is itself tentative, depending heavily on the relative importance to cable systems of STV income, general subscriber income, and advertiser income. Until these proportions emerge more clearly our textual propositions seem sound.

channels; or (3) further adjustments by broadcasters in TV content and regionalized service. At issue, finally, is the continued policy on cross ownership to forestall that response to cable's inroads, especially whether that policy should be modified to permit cross ownership in out-of-town markets only.

Sloan had in its possession data to answer some of these questions. *Prospects and Policies for CATV* constitutes a major source of such data, granted its controversial character. A fuller divulgence of Sloan's evaluation of that critical document and its bearing on the questions just outlined, could have enlightened the discussion of related issues.¹⁵

As matters stand, neither Sloan nor the FCC considered cable's interim impact on *both* broadcast *and* cable rents and service. However it will be a long time before we have penetration of even 30 percent, and effects on both modes were therefore in bad need of review.

Independent economic studies predict cable's main adverse impact will not be on independent UHF stations at large, or in the largest markets (they may actually be *helped*). The greatest revenue losses (over 50 percent) will be suffered by VHF affiliates in the smallest markets.¹⁶ However, VHF affiliates are elsewhere not the area of greatest expected attrition, so it is hard to understand why unlimited signal importation should now be opposed for fear of eliminating the network news and public affairs function. Furthermore, this opposition ignores cable's future structural adjustments, possible large-scale program economies via satellite/microwave interconnection, and the emergence of new program services. It also ignores cable's possible extension of the audiences of VHF independents in the big cities, or otherwise.¹⁷

Was this new case for the ban on importation just a power play by the TV networks, or by an FCC wed to the formal facade of an obsolete TV Allocation Plan? Can we rightly ask for something back in public access or educational channels for entry privileges we had no right to take away in the first place? True the importation rights are now very valuable, and without them cable penetration and the cable surplus will be much lower. But if the importation ban was supposed to preserve the TV Allocation Plan and its priorities (with-

¹⁵ Without such divulgence and discussion one is left with the intriguing matter of that portion of the McGowan-Noll-Peck paper (*Prospects and Policies for CATV*) omitted from the Final Report, but included in the unabridged version distributed on request. The whole concept of cable surplus, its determinants and measurements, its relevance to evaluating policy alternatives, are absolutely crucial to the *Sloan Report*. With no explanation of why this was cut out, and no airing of the professional disagreements over the M-N-P assumptions and conclusions, we are at best left with an ambiguous situation. At the very least, those documents on which Sloan based its judgments should have been divulged. See further discussion below, footnotes 24, 25, 27-30, 41, and associated text.

¹⁶ Johnson. *Cable Television and the Question of Protecting Local Broadcasting*, Rand Memo R-595-MF, Oct. 1970, pp. 5-6; and Park [7], pp. 72-73.

¹⁷ It is well to recall that the network-affiliate relation originally emerged from proscriptions against superpower, group and chain ownership. But the eventual national network arrangements gave us large-scale programming economies, and the wherewithal for internally subsidized news and public affairs. With sufficient market access, some cable program capabilities not now available may in fact emerge, so that those who would retain the ban on signal importation have a double burden of proof.

out which there would be no need for any ban), who is being protected for what? How much channel capacity for whom is at issue, and how is it to be used? Sloan unfortunately neglected these matters.

Unquestionably, if importation rights facilitate cable penetration and an economic surplus, the power of any regulator who awards them over that surplus and its eventual disposal is strengthened immeasurably. Accordingly, if regulators want to "brake" cable growth, candor requires that they make their premises more explicit and more fully documented. Otherwise the new network argument becomes a hollow facade to defend broadcast profits and rents, no more palatable than the older arguments still cited to safeguard the TV Allocation priorities. In expounding its affirmative position on signal importation and cable growth, then, Sloan should have made more explicit the original ban's several tacit objectives and rejected them head-on.

Another omission was failing to consider, if only to rebut, some device to activate the UHF (via subsidy) and thus to secure in service and diversity what we now look to cable for. An excise tax on TV set owners might provide revenues to activate the UHF band commensurate with the subscriber's growing cable investment and with the volume of cross subsidies needed for public access channels. Cable subscribers are now "buying" twelve channels at a price which may well exceed what it would cost them to subscribe to ten off-air VHF/UHF channels, and the Sloan Commission should at least have analyzed this explicitly.

The problem is that really to know we would have to institute an extensive pay-TV system in the UHF, or to activate idle UHF channels with an adequate subsidy. In either case there would be difficult problems of public funding, public ownership, and new organizational arrangements, whereas cable provides a possibly more acceptable route to achieve the same end of multi-channel diversity.

Nevertheless, broadcaster power to stop cable may have been exaggerated. It is at least harder to stop a new technology than to oppose modifications of an established one, such as low-powered UHF stations, narrow VHF spacing, or license taxes to activate an idle UHF. Even if more costly for cable subscribers to buy twelve channels directly, to do it via the UHF requires extensive institutional revamping, and this should have been faced squarely.

□ **Released spectrum resources as a factor in cable growth.** Aside from the FCC's protective stance, there is still another factor that could influence the degree to and speed at which cable realizes its potential. I have in mind the spectrum released by cable, and its relevance to phasing in the new technology over time.

Although politically understandable, it is disconcerting that the Sloan Commission and the FCC should both ignore the economic implications of cable's substitutability for radio frequencies.¹⁸

¹⁸ Except for passing comments (Sloan Report, pp. 47-48) and a sketchy background paper it never released (Passell, *Allocation of the Radio Spectrum*, March 1971, 14 pp.), Sloan appears to have given little attention to spectrum. The potential significance of reduced frequencies for phasing in the new cable technology over time is not even mentioned among the "important issues" the *Sloan Report* "does not cover or merely alludes to in passing" (Report, p. 168). The sole major reference to spectrum relates to the regulatory constraints hitherto

Surely TV's costless access to spectrum has been a serious source of broadcast regulatory problems. Had broadcasters paid the full value of their occupancy rights, the need to exact a compensating cross subsidy for public service would be far less pressing. But spectrum charges on access rights have thus far been impossible to institute, so that the resultant spectrum scarcities underscore all the more the need for a serious analysis of cable's implications.

Cable inroads into broadcast markets could induce a growing regionalization of TV service, whereas the plentitude of cable signals will weaken the case for indefinite retention of the underutilized UHF band. The opportunity costs of unused spectrum are considerable¹⁹ and represent potential resources to help phase in cable by: (1) compensating the networks for any serious damage to news and public affairs; (2) subsidizing the remaining UHF stations; (3) extending cable to sparse rural areas; and (4) installing broadband circuitry. In particular, the introduction of a switched interconnected cable system could go far to replace the national TV network system and hence to release still more UHF and VHF spectrum on that count too.

The most "way out" ramifications of released spectrum, then, are the resources it represents relative to the cost of laying a nationwide cable grid, or broadband circuitry generally. Sloan's willingness to explore Bell's possible role in laying out such a grid²⁰ was never matched by any comparable attempt to weigh the value of released frequencies in phasing in a national cable network. To say that the broadcaster's political sensitivities ruled out any such analysis begs the question. The FCC's desire for a working compromise (as in the consensus agreement) may explain *its* reluctance, but Sloan asserted a greater responsiveness to public than to private interests, and this would lead us to expect better.²¹

In matters of spectrum economy and regionalized service, in any case, the Justice Department went much farther than Sloan. Furthermore, though TV's free access to spectrum may not be legally relevant to the cable copyright issue, Sloan should at least have reviewed its analytic implications for "unfair" competition between the two media.²²

Sloan's failure to examine spectrum issues more closely is part of

imposed by spectrum scarcity and which cable may now relax (*Report*, pp. 16-22, 42-46), and the technical interference problems associated with cable channel usage, their technical solutions, and their cost and regulatory ramifications. (*Present and Probable CATV/Broadband Communications Technology*, Sloan Paper by J. Ward, Jan. 5, 1972.)

¹⁹ See generally, Levin [6], chs. 5, 12.

²⁰ Minutes of Third Meeting of Sloan Commission, Oct. 15-16, 1970.

²¹ *Sloan Report*, pp. 80-83, 143-447. The closest Sloan came was in (1) noting the possible regionalization of TV service in response to cable's inroads (and presumably the reduced spectrum needs this implies); (2) prescribing speedy increases of cable capacity whenever demand exceeds supply, proscribing any indefinite nonuse of mandated cable channels, in both cases presumably to forestall the kind of problems which spectrum scarcity has long posed in broadcasting.

²² Granted that the frequent sales of TV stations may by now have imposed *de facto* charges on many if not all TV licensees for the spectrum they occupy. But all this implies is that we must check and compare comparative hardware costs for the two transmission modes. See Comments of U. S. Dept. of Justice in FCC Docket No. 18397-A, Dec. 7, 1970, pp. 4-8.

a related lacking, viz., a neglect of what we earlier suggested could be a protracted interim coexistence of cable and broadcasting. Sloan also ignored the possible accommodations of the two media in programming, rates, etc. The discipline imposed by TV broadcasting on cable rates is noted by those who oppose premature rate regulation for cable subscribers, or for cable channel lessees. But Sloan and the FCC paid virtually no attention to the *long-run* program adjustments,²³ albeit in the short-term cable may be more an auxiliary device which simply extends the existing broadcast service.

■ For present purposes, the “cable surplus” is simply “profits in excess of those necessary to induce CATV investment.” To say that any effective surplus is *in fact* regulated out of existence, or that the local cable system is not enough of a natural monopoly to generate very much surplus, really misses the point. The cable surplus is a potential which may or may not materialize. And its size will depend on what the regulators do to subscribers rates, lease charges, non-exclusive franchises, access to the municipality’s subterranean ducts, and a host of other factors.²⁴ It will depend also on how successful the many claimants of any potential surplus are in wresting their share.²⁵

Once cable comes of age an overriding question will be how it can help us better do what we have long tried but failed to do in broadcasting. Of the several critical areas Sloan probed, two will be considered here.

□ **Mandatory cross subsidies.** The hallmark of broadcast regulation is still to recover a diversity and quality of service greater than would otherwise flow from a market without direct prices per program or per channel, where entry is impeded by spectrum scarcities which *ceteris paribus* operate to raise duplication levels, where advertising ratings methodology is at best imperfect, and where merit programs require subsidies of some sort, external or internal. From the very outset the Commission has viewed industry health and financial resources as necessary for such meritorious, but sometimes unprofitable, service as news, public affairs, special events, documentaries, etc.

Traditionally faced with entry barriers grounded on spectrum scarcities, the Commission has made necessity the mother of invention, imposing a *de facto* principle of qualified profit maximization on all licensees to provide, through voluntary internal cross subsidies, what under other technical/regulatory frameworks might have been done more directly. Sometimes compulsory cross subsidies have been proposed, as in a prime time rule to handle public affairs including Presidential campaign debates. The fairness doctrine more generally,

²³ The *Sloan Report* largely ignores this issue. (See *Report*, p. 81, and chapter 7 generally). Perhaps the most suggestive approach for analysis of potential long-run adjustments appears in Johnson, *Future of Cable Television: Some Problems of Federal Regulation*, Rand Memo RM-6199-FF, Jan. 1970, pp. 66–72. Also more generally, see Levin [4], chs. 5–6.

²⁴ See further discussion above, footnotes 14–15, and below, footnotes 27–30, 41, and associated text.

²⁵ See *Prospects and Policies for CATV*, p. 1.

3. The cable surplus and its disposal

and Section 315 in particular, contain similar internal subsidy elements. In all such cases one encounters the twin problems of vague, elusive standards and direct federal intrusions.

How will cable respond to this linkage of economic surplus and public service? Whatever else one presumes about cable's alleged natural monopoly status at the local level,²⁶ profit rates will clearly depend on the permissible (or market-determined) level of subscriber charges, lease rentals, and possible advertiser income. These will in turn depend upon effective cable penetration, itself heavily dependent on the availability of attractive programming.

Accordingly, distant signal importation, combined with a monopolistic transmission mode, could force us to choose between a rate regulation which could deter the inflow of needed venture capital; tolerating supernormal profits; or siphoning some of them off as mandatory cross subsidies for special purpose channels. Sloan mainly elected the third option,²⁷ but how much was this a "forward" step from broadcast regulatory doctrine and administrative practice?

Public service responsibilities of broadcast licensees are sometimes alleged to pose First Amendment problems and, however invalid the criticism, there is little question that structural safeguards are a preferable approach to diversity. But the problem today is that multi-commercial broadcast outlets normally fall well short of the mark, while local community outlets remain an unrealized paper promise. Only public television operates as an alternative of potential significance. In contrast, cable offers sufficient capacity for a truly diversified service and enough channels to permit at least some to be "mandated" for special purposes.²⁸ Such mandatory usage offers a potent technique to convert economic rents into public service, all without risking the regulation of specific program contents.

Nevertheless, Sloan needlessly weakened its position by failing to

²⁶ Sloan was ambivalent about the local monopoly character of cable systems. Compare affirmative statement in *Sloan Report* at p. 5 with comments at p. 139, where Sloan foresees many markets as able to support competing cable franchises. Nevertheless, the M-N-P "surplus" presumes a local natural monopoly grounded on scale economies and the avoidance of physical duplication. This is presumed also in several Justice Department filings: Comments in Docket No. 18397-A, Dec. 7, 1970, pp. 17-19, and in Docket No. 18397, Sept. 5, 1969, pp. 5-11. Effective competition for access to the same household even in markets supporting multiple systems does require more detailed analysis than so far forthcoming. Some of the most critical issues are touched in a very interesting paper by the Stanford Institute for Public Policy Analysis cited below, note 29.

²⁷ See *Sloan Report*, ch. 12, and *Prospects and Policies for CATV*, pp. 49-63.

²⁸ "Mandated" channels designate the uses to which some proportion of given channels shall be put. Sloan proposed that half of a 20-channel system should be "mandated," with six channels for network and local signals, two channels for government and nongovernment services, and one channel for experimental and educational. Of the remaining ten channels, eight would be open for leasing on a first come, first served basis, to any bidder, profit or nonprofit, public or private, political groups, etc. Franchises were to be awarded to applicants who promised the lowest rates and highest quality services, with initial specifications of standards and services and rates at which provided, and subsequent second bids for some approved prototype package of technical-service requirements. The low bidder would win on the second round. Hence the second requirement was for internal subsidization of some services: the charging of nondiscriminatory subscriber fees, the bringing of cable into rural areas, the provision of low-cost head-end facilities for public access channels, the installation of capacity initially far beyond the system's needs. Sloan much prefers such internal subsidies to any special tax on cable which it fears must stunt its growth. (*Sloan Report*, pp. 141-146.)

deal head-on and in greater depth with a number of perplexing corollary issues.

(1) *Cable surplus*

- * If rate regulation were really rigorous, need there be any economic rents at all left to dispose of via mandatory cross subsidies? Aside from rate regulation, any projected surplus could be sharply reduced by awarding nonexclusive cable franchises and by guaranteeing access to the community's commonly owned ducts *and* to the subscriber's "black box" (the interface between TV receiver and cable installation).²⁹ Would automatic capacity increases (whenever demand exceeds supply), not further buttress these arrangements?³⁰

(2) *Excess cable capacity*

With automatic increases in capacity would all channels always be leased commercially anyway? If not, why mandate special purpose usage? Mandatory free access to public channels may raise costs little today, but this could change should channel demand rise markedly. If advertiser income becomes important in cable, systems owners will admittedly have to worry about fractionated audiences and competition from other media. But if subscriber income predominates, the systems would have greater economic incentive to diversify their package on the merits. In that case, wouldn't the urgency of mandatory cross subsidy diminish?

(3) *Merit programs*

Are cross subsidies under any condition needed even for merit service on the cable? With per program or per channel pay systems, may not low-cost satellite interconnection of functionally related specialized audiences and sheer channel capacity, make viable minority interest or merit programs that are not now viable? The opportunity cost of catering to minority tastes on a marginal cable channel would be much lower than under broadcasting.³¹ So that if channel abundance, pay-TV, and low-cost interconnection enable us to piece together easily audiences of 600,000 for special topics, a merit good requiring external (or internal) subsidy today, could well

²⁹ Barton, Dunn, Parker and Rosse, *Nondiscriminatory Access to Cable Television Channels*, Stanford Institute for Public Policy Analysis, May 1972, pp. 37-40. Sloan's judgment that many markets can support multiple systems due to limited scale economies underscores the potentiality of nonexclusive competitive franchises even though it never draws that specific conclusion. But see *Sloan Report*, p. 139.

³⁰ The FCC's sustained efforts to insulate broadcasting against severe cable inroads may also operate inadvertently to check cable subscriber rates and advertiser rates. But the protection of broadcasting is neither an efficient nor a deliberate attempt to further such an end.

³¹ With unlimited channel capacity, the marginal cost of any *n*th channel will approach zero and cable operators will therefore want to put on as much service on as many channels as they can. This makes for a more attractive system. True, cable advertisers may bar fractionation of the audience at any given time, and will naturally not pay as much for a smaller audience. But substitution of public service is not all-or-nothing-at-all. One does not deprive customers of anything by running a public service program. If indeed the package is made more attractive one can even raise fees against the commercial users and advertisers.

become marketable tomorrow.³² Nevertheless, may it not be cheaper to reach the 600,000 viewers within a *single* market of several million even if four-fifths of the larger audience do not watch? Will this not depend on the emergence of brokers to piece the audiences together in ways that minimize the risk of defection to attractive counter-alternatives?

(4) *Importation rights*

Rights to import distant signals are very valuable today and Sloan, like the FCC, wanted something back from their recipients. But wouldn't this simply recover something improperly withdrawn in the first place? By first banning importation and then relaxing the ban at the price of mandatory cross subsidy, may the FCC not have recreated (and Sloan endorsed) the old broadcasting dilemma all over again?

If signal importation does threaten broadcast structure and service it may of course make good policy sense to insure a compensating program diversity on *cable*. But is mandatory cross subsidy the best way to proceed?³³ Or will vested interests later block advances in cable technology in the name of FCC's regulatory priorities? Will subscriber rates, set artificially high to make cross subsidy possible, not impede cable growth simply because some intended beneficiaries will just not subscribe?

This is no place to try to answer such questions, but Sloan should have handled them more explicitly. Clearly, if one really wants diversity and merit programs *beyond* those likely to materialize in an unregulated market, one must subsidize in *some* way. General appropriations, nice in theory, pose serious practical problems in broadcasting and it may well be better to use an imperfect mechanism because a theoretically more ideal one is impractical or politically unrealistic. Granted that if one sanctions internal subsidies in cable, as elsewhere, one must specify the performance dimensions being subsidized, and at least try to quantify the ultimate costs and benefits.

□ **Relations with public television.** The FCC provided spectrum for 242 public television stations in 1952, eight years before Congress gave the states matching grants-in-aid to build physical facilities, and fifteen years before creating the Corporation for Public Broadcasting. Yet fewer than that number of stations are actually licensed today, while the spectrum reserved for hundreds more still lies idle. Since 1952, then, the ETV spectrum reservation has obviously imposed high opportunity costs on the nation, costs which might never have occurred had there been healthier respect for the spectrum's economic value to next-best users in commercial broadcasting and elsewhere.³⁴

As conceived by Sloan, does cable offer any significant step forward for public television? In some ways, yes. As soon as commercial cable systems are built, there *ipso facto* become available extra channels for local educators, and for the mandatory carriage of local PTV stations. The long step in broadcasting between spectrum al-

³² *Impact of Satellites on Cable Communications* (Sloan Paper by A. Chayes, May 1971), pp. 15-24.

³³ See Posner, [8, 9].

³⁴ See [6], pp. 351-356.

location and physical facilities construction simply has no counterpart in cable, and Sloan could have made more of this fact.

True, Sloan did proscribe indefinite nonuse of any mandated channels, presumably including those set aside for public television.³⁵ But when one thinks of comparable warnings during the long nonuse of reserved educational spectrum, this may turn out to be more a paper promise than a real one. Sloan should at least have drawn attention to the rather different problems posed by unused *spectrum* (which requires hardware, program outlays and time to activate), and unused *cable facilities* which at least require no additional hardware.

The program resource component of public television, on the other hand, has so far posed hurdles in cable no smaller than those in broadcasting. One recurrent "solution" has been a Public Dividend Plan wherein commercial cable systems would pay 5 percent of their gross receipts to CPB for broadcast use.³⁶ This is reminiscent of an earlier Ford Satellite Plan (1966) which proposed to fund public programming and public channels from part of the cost-savings due to satellite technology, itself a product of sizable public investment in rocketry and orbital components.

In cable a different rationale was clearly needed, but never articulated by the Sloan Commission or the FCC. One version might be that cable inroads into commercial TV will reduce the latter's wherewithal for merit programs or, indeed, public television's own support base. Therefore commercial cable might be asked to make good any such impairment. Yet this argument is often improperly mouthed at the very time the FCC *insulates* broadcasting against cable. And the argument does not consider commercial TV's potential competitive response to cable, its associated contribution to diversity and, via pay-TV and low-cost nationwide interconnection of functionally specialized audiences, to merit programs as such.

A further objection to the Public Dividend Plan rests on its predicted deterrence of new cable entrants. Some observers fear it could kill the goose that lays the golden egg, and others fear for centralized control by CPB. Still others see the new cable technology, long before its monopoly rents accrue, as intolerably burdened to support a nonprofit broadcast alternative even though public broadcasting is no "chosen instrument." To protect public broadcasting before something else is ready to take its place, they say we may virtually stifle that "something else."³⁷

Neglected further in the Sloan Report and papers was a possible gross receipts charge on the whole cable system, paid by all lessees and earmarked for a nonprofit program subsidiary. Here would be a local (not centrally controlled) program resource, to give substance to the paper promises of local educational and public access channels. Perhaps the closest Sloan came to this conception was in rejecting all cross ownership limits involving public TV, cross ownership which

³⁵ *Sloan Report*, p. 144.

³⁶ See generally, "Second Further Notice of Proposed Rulemaking" in FCC Docket No. 18397-A, July 1, 1970, pp. 4-5; Comments of Corporation for Public Broadcasting in FCC Docket No. 18397-A, Dec. 7, 1970, pp. 4-7. See also Johnson, *Cable Television and the Question of Protecting Local Broadcasting*, Rand Memo R-595-MF Oct. 1970, pp. 24-25.

³⁷ Jones, *Regulation of Cable Television in New York State*, Dec. 1970, pp. 201-204; Comments of U. S. Dept of Justice in FCC Docket No. 18397-A, Dec. 7, 1970, pp. 13-14.

might presumably yield money to support educational and public access channels, or local originations more generally.³⁸

One final issue relates to the ultimate cost of free or reduced-rate microwave or satellite channels to interconnect local cable systems. However funded today, public broadcasting is a clear alternative to negotiated internal subsidization of merit service funded out of VHF monopoly rents. So that while preferred access to spectrum by educators (with long interim nonuse) may pose high opportunity costs, so does the negotiated cross subsidy approach.³⁹ Sloan's treatment of public channels and mandatory cross subsidy would have benefited from analysis of this point, and from a more visible evaluation of the preferential pricing of cable and broadcast interconnection.

□ **Public access channels.** The thorough balanced treatment of controversy, long a principal FCC touchstone, figures in the "bluebook" doctrine of 1946 as applied in initial licensing, license renewals, and recent renewal contests; in the call for paid or free access for "ad editorials" on public issues; and in the call for more news, documentaries, special events and other public affairs programs. Adequacy in this area is normally viewed as requiring the FCC to goad broadcast licensees into internal subsidization.

The problem is how to define and administer practical, meaningful program standards under the fairness doctrine, and how to determine when free or paid time is appropriate. It is hard at best to balance between the disincentives of extreme uncertainty due to renewal contests, and *de facto* perpetual enfranchisement, however insensitive the broadcaster to minority claimants on scarce broadcast time.

Sloan saw in public access channels a major breakthrough to resolve such problems.⁴⁰ Located within well-defined neighborhoods with ethnic problems, these channels would hopefully enable the groups therein to speak *gratis* to each other, to the inner city, and to the general public.

Sloan's safety valve (and the FCC's) seems twofold: (1) sufficient channels to devote one exclusively to public discussions at the grass roots level and, (2) sufficient economic surplus for any needed cross subsidy to afford the eligible participants free access.⁴¹ The objective is to air the widest range of views with no need for the FCC to (1) review directly the program content itself or (2) negotiate by raised eyebrow for internal or external subsidy of particular program types. The actual program content will presumably emerge from an interaction of the excluded groups. The cross subsidy will provide the channel—its physical transmission hardware and studios—not the transmission time, which is abundant because the multiplicity of channels will keep the associated opportunity costs near zero. What excluded minority claimants now demand of broadcasters in paid or free access, or in the right to air countercommercials, they can

³⁸ *Sloan Report*, p. 141.

³⁹ See generally [6], pp. 167–171, 351–356.

⁴⁰ *Sloan Report*, ch. 11, and Appendix C; FCC Cable Rules, Title 47, ch. 1, Sec. III, paras. 117–148.

⁴¹ See *Prospects and Policies for CATV*, pp. 49–63. For a well-documented contrary view compare Comanor and Mitchell [2].

presumably get in part at least by direct interaction on their cable system's public access channel.⁴²

Or can they? Who will be listening? Without equal time guarantees, how much of a safety valve will the channels actually provide? If an "offending" program reaches a very large audience, and the "reply" a very small one, the fairness doctrine will still be needed, as Sloan quite properly notes. However, Sloan would have us lift that doctrine in cable as often as possible to forestall the timidity it now imposes on broadcasters.

Yet in broadcasting, equal time requirements have often been circumvented by *paid* coverage of the issues, where paid access is obviously not free access. Analogously, the mere provision of free public access cable channels could be a pyrrhic victory if the main intended users lack adequate resources or expertise to utilize them. Neither Sloan nor the FCC has given this crucial matter enough attention. Since only a system's first public access channel is to be free, moreover, what about the price of a second one? And how should one handle these additional problems: the danger of preclusive rates; the need for equitable scheduling and time sharing; the need for right-to-reply rules for overt attacks and defamation; the need for new legislation on liability for obscenity and incitement to riot, etc.? Will the many rules needed to minimize these risks themselves act to stultify use of the public access channels?⁴³

Perhaps one cannot expect a full future blueprint, but Sloan should at least have considered whether, in the absence of excess channel capacity, cable owners will continue to screen on a first come, first served basis. The Commission also neglected considering what the chances for automatic increases of channel capacity will be whenever demand exceeds supply, how quickly lump-sum or incremental costs will rise when one reaches certain key thresholds in capacity (above 40 channels), and finally, if capacity increments are lumpy or sticky, whether cable owners will not have to ration possible queues for public access, ETV, or other leased channels—and with what result.

How also would Sloan have us avoid repetition of innumerable broadcast regulatory myths—the local live myth, the UHF myth, the ETV myth? Paper promises are not enough. Yet mandatory special purpose channels without the resources to use them may one day impose sizable opportunity costs. Clearly the earmarking of TV spectrum to sustain an unrealized local service priority (or an uneconomic UHF or ETV structure) has already imposed very comparable opportunity costs. In contrast, tapping the cable users to contribute to a nonprofit programming subsidiary might, as noted, reduce the cost of underutilized public access or educational channels.

Nevertheless, cable may not diversify its service if advertising is permitted and if one person owned all the channels on any system. Advertising sponsors are probably best won by transmitting the same program on all channels simultaneously. Therefore, the answer

⁴² The broadcast access rights analogous to those created by public access channels on cable are set forth in *Red Lion Broadcasting Co. v. FCC* 395 US 367 (1969); *Banzhaf v. FCC*, 405 F. 2d 1082 (D.C. Cir. 1968); *Business Executives Move for Vietnam Peace* (D.C. Cir. 1971). See generally, Barron [1].

⁴³ For a suggestive proposal on how to avoid this danger by creating a local nonprofit programming entity fed by a 5-percent surcharge on all commercial cable users see H. Frederiksen [3], pp. 17–28.

is probably to limit the operator's ownership rights *and* to follow Sloan in the mandating of selected channel usage. At the very least, mandatory cross subsidy of special purpose channels will guard against any profit-maximizing channel usage which lessens diversity.

4. Organization, ownership, and control

■ In broadcasting, the main seat of power lies with three national TV network organizations. Recurrent attempts to contain that power still fail to reduce prime time clearances over the hundreds of affiliated stations to less than 90 percent and, foreclosed access even now is sufficient to elicit a Justice Department antitrust suit.⁴⁴

Can cable be structured so as to prevent recurrence of these and other concentration problems? The Sloan Commission addressed itself to this question at several points including a discussion of the policy on cross-ownership and the possibilities of common carrier organization. My sense is that apparent advances in the former may not in fact materialize, whereas an understandable reluctance to adopt the latter outright may be an irrevocable error.

□ **Cross ownership of cable systems.** At first glance, Sloan's handling of ownership and control constitutes an advance over the broadcast situation. First, the national TV networks are entirely excluded from cable. This maximizes the likelihood of competition between broadcast and cable television, and of unencumbered vertical access by independent nonnetwork programmers.⁴⁵ No analogous action has been possible in TV network ownership of broadcast stations after the fact, despite the cogent *ex ante* case sometimes heard.

Second, no cable company may hold systems with more than a specified number of cable subscribers nationally (10 percent of all households within the system's reach at any time).⁴⁶ This proposal approximates one sometimes heard to limit TV group ownership by aggregate station circulation, not as at present by the mere number of stations.

Third, the cable holdings of co-located TV stations will be limited to 10 percent of the SMSA's households, and 40 percent nationally.⁴⁷ Once again this is more explicit than the bald application of broadcasting's so-called duopoly rule, wherein under only one AM, one FM, and one TV station can be jointly held by any licensee whatever the size of each station's circulation.

Fourth, in addition to the mandating of public access channels, Sloan proposed that local franchising authorities should favor non-profit (or profit-making) authorities in communities with special social or ethnic needs.⁴⁸ This is reminiscent of preferred access to spectrum by public television, though it is less clear how (or if) profit-making entities will efficiently deal with such problems.

Fifth, public television stations are, as noted earlier, subject to no limit at all on their cable holdings.⁴⁹

⁴⁴ See *U.S. vs CBS, NBC, ABC*, USDC, Central Distr. of Cal., Civil Actions No. 72-820; No. 72-819R5R; No. 72-821LTL, April 14, 1972.

⁴⁵ *Sloan Report*, p. 140.

⁴⁶ *Ibid.*, pp. 140-41.

⁴⁷ *Ibid.*, p. 140.

⁴⁸ *Ibid.*, p. 162.

⁴⁹ *Ibid.*, p. 141.

Sixth, Sloan opposed complete separation of ownership from programming at this early stage of cable growth,⁵⁰ a question to which we return in our concluding section.

Seventh, newspapers (like TV stations) are limited to cable holdings that reach no more than 10 percent of local market households and no more than 40 percent nationally.⁵¹ However, a number of serious loopholes question whether the ceilings will be effective, and these pose problems of sufficient import to warrant special treatment next.

Newspaper applicants for TV broadcast stations are still screened under procedures first formulated in 1941. When two or more candidates apply for the same outlet, other things equal, the license is supposed to go to the nonnewspaper or the candidate without other media affiliations. Actually, nonnewspapers have been preferred to equally qualified newspapers to promote diversity of expression. But newspapers with satisfactory legal, technical, and financial qualifications have never been denied licenses only on grounds of newspaper ownership, when they were sole applicants. Finally, when other things are not equal, newspapers have been preferred to nonnewspapers. Sometimes the newspaper wins on grounds of superior resources and experience. In other cases, civic-minded, locally resident newspapers, aware of their community needs, have won even when they would thereby operate the only station in town. Still others have won because they showed greater initiative in organizing local talent and formulating program plans, and because they resided in communities with little or no service compared to the communities the rival candidates proposed to serve.

At first glance, Sloan's explicit ceilings on cross ownership may represent an advance over current FCC practice. The multi-channel cable structure is also a further safeguard, provided no owner may program for more than one or two channels out of twenty or more. And unencumbered vertical access is still another safeguard. For the leased channels (half of the total minus two) are open to all comers, except for two on which the owner may program.⁵²

But how *effective* will these safeguards be in light of several neglected loopholes? Where no nonnewspaper applicant seeks a license, Sloan's quotas would be waived.⁵³ Judging from the experience in broadcasting, this could be an almost irremediable loophole. Surely the most frequent entrants into broadcasting have been sole applicants⁵⁴ who, once installed, are virtually impossible to dislodge. Cable owners are not in any case explicitly prohibited from program transmission on any *idle* lease channels (theoretically open to outsiders, but not in fact used by them). Nor is the price at which lessees are expected to enter entirely clear. Nor whether an owner can be kept from exclusionary pricing to protect his own (two)

⁵⁰ *Ibid.*, p. 148.

⁵¹ *Ibid.*, p. 140.

⁵² The potential safeguards offered by the multi-channel cable structure is directly analogous to that offered by the multi-station structure in TV broadcasting. In the broadcast case, any potential anti-competitive abuse of TV group ownership or conglomerate ownership is forestalled by the effective intra-market TV structure. The same presumably holds true in single (or multi-) system cable markets. (See Levin [5], pp. 794-96, 801-02, and Appendix A, Table 2).

⁵³ *Sloan Report*, p. 141.

⁵⁴ See Levin [4].

channels, or to deter potential lessees from entering, thus enabling him to program on *more* than two channels.

Finally, there is no stated limit on the number of leases anyone can hold on *all* cable systems (though no one may hold more than two leases on any *single* system).⁵⁵ Therefore, newspapers could lease two channels on all systems in the country, and so could each TV network, with no limit on total subscribers.

Sloan might have used the broadcast experience with cross ownership as a more explicit benchmark for cable policy. It could have distinguished better between potential and actual abuse, explicit editorial bias and a more subtle slanting. Even "content analysis" is methodologically limited, so we must distinguish further between structural safeguards and direct policing—the former being a key alternative to the problems posed by excessive administrative control, and the stultification it imposes.⁵⁶

Affirmative new concepts must be instituted. Why run the risks of joint ownership if separate ownership does not actually reduce diversity? There are after all other sources of investment and funding, so that a painful restructuring is by no means the only way to proceed; preventive action is also important.

To be sure, the lease channels of even a newspaper-owned cable system provide structural safeguards of a sort. But are such channels comparable to the existence of several separate cable systems in the same community⁵⁷ or to separately owned TV broadcast stations? Would the multi-channel cable structure at least work satisfactorily if the newspaper were a pure common carrier? Or will the need for even a newspaper common carrier owner to schedule programs with an eye on advertiser needs (and hence to affect content) pose the traditional threats to diversity? How many channels should a newspaper be permitted to lease on any co-located cable system, whether or not it *owns* the system? What potential conflicts of interest may arise in joint advertising rates, in cable development and in exclusionary pricing of leases? One misses a careful review of these issues in Sloan's cross ownership discussion.

Stated differently, will the mere prescription of nondiscriminatory vertical access to newspaper-leased channels suffice to safeguard *non-newspaper* access? And is this likely unless nonexclusive franchises are instituted, and access to ducts by alternate cable systems guaranteed, to keep lease prices in line?⁵⁸ What one really wants is a variety of policy options, ranging from no cross ownership to no limits at all, with the actual ceilings set according to the effective number of media voices, alternative sources of capital, number of

⁵⁵ *Sloan Report*, p. 143.

⁵⁶ See Levin, *The Policy on Joint Ownership of Newspapers and Television Stations: Some Assumptions, Objectives and Effects*, Statement in FCC Docket No. 18110, April 1971, p. 1-31, and in particular, Reply Comments, August 1971, pp. 1-8. More generally see also Levin [4], Chapter 3.

⁵⁷ The multi-channel cable structure should operate to safeguard nonnewspaper access so long as newspaper owners of a cable system are permitted to operate only two channels, or, if on some unused leasable channels too, only until a nonnewspaper lessee applies. (*Sloan Report*, pp. 142-43). Strictly speaking, however, account must be taken here of the possible divorcement of systems hardware ownership from programming and transmission rights. (See below, pp. 361-364.)

⁵⁸ See citations note 29 *supra*.

subfranchisees, degrees of access guaranteed, number of channels on the system, credibility of new entry, character of alternative entrants, their potential conflicts of interest, etc.

But in Sloan one finds instead a compromise position where (1) scale economies are such that markets can support competing cable systems, each well above the minimum economic size; (2) specific quantitative limits are imposed on cross ownership; (3) unlimited cross ownership is permitted out-of-town; and (4) cross ownership by public television, radio, and nonnewspaper publishers, is also left unchecked.⁵⁹

It gives pause, moreover, that Sloan heard ANPA witnesses but no independent critic from the other side.⁶⁰ Nor did newspaper representatives see any substitute for outright cable ownership by *them* to insure against the “inevitable” editorial bias of the non-newspaper owner.⁶¹ This is no small irony in view of ANPA’s angry denials whenever its members are criticized for “inevitable” *newspaper* bias. Finally, Sloan’s unwillingness to exclude cable owners entirely from programming should be recognized here for what it really is—another argument in support of a bolder, not a weaker, position on cross ownership.

The merits of cross ownership must of course not be lost from sight: the inflow of expertise, talents, skills, capital; insurance against powerful newspaper opposition to cable growth; the equity of letting in those whom cable most threatens. But it is by no means clear why these benefits could not be derived adequately from a scheme which limits cross ownership exclusively to out-of-town markets.⁶² Or why, if newspaper and TV cross ownership with cable were handled rigorously, cable links with public television or radio should remain entirely unregulated.

□ **A common carrier approach?** If the Sloan proposals on economic concentration are plagued with their own limitations, is a common carrier approach any better? Can it keep cable from restricting access and output, incentives to the contrary notwithstanding? Are the structural safeguards just discussed, enough—the limits on cross ownership, the exclusion of TV networks from cable, the mandating of public access and educational channels, and of leased channels generally? Or is something more needed—the outright divorcement of systems hardware ownership from programming?

In broadcasting, regulatory problems often arise from the fact that licensees own their own transmitters and studios, prepare their own programming, or, more likely, get it from three national TV networks. Problems arise also because the networks own key stations and interconnect hundreds of affiliates across the country; produce,

⁵⁹ *Sloan Report*, pp. 139–41.

⁶⁰ *Sloan Report*, Appendix F; Minutes of Sixth Meeting of Sloan Commission, Jan. 21–22, 1971.

⁶¹ *Ibid.*

⁶² Prof. Stephen R. Barnett has argued convincingly in favor of limiting all newspaper-cable cross ownership to out-of-town markets where the two enterprises are not co-located. See comments of Stephen R. Barnett in FCC Docket No. 18397, April 1, 1969, pp. 194–99. In contrast Sloan’s passing reference to such nonlocal cross ownership (*Sloan Report*, p. 140) slighted a potentially important way to retain the merits of cross ownership without tolerating its dangers. (*Ibid.*, pp. 137–39).

buy and distribute programs for them (and may even create their own domestic satellite to do so); and secure the vital national advertising funds. Although cable need not be organized the same way, the danger of "regulatory hangover" is considerable.

The three-year broadcast license term has had little practical significance. *Pro forma* renewals make licenses with even three-year limits *de facto* property rights except, of course, where licenses are challenged at renewal, and where judicial endorsement of the challengers' rights now poses more tangible risks for incumbents.⁶³ The dilemma is how to choose properly between demoralizing uncertainties and disincentive effects on one hand, and *de facto* property rights and lax enforcement on the other.

Cable is in a good position to avoid this dilemma, something Sloan should have pointed out squarely even if it did not approve. The key is to divorce the ownership of physical assets from program production, and even from scheduling and interconnection. Such a break has already occurred in the British commercial television industry. Because program contractors there have no hardware investment, disenfranchisement imposes far less duress than in the United States. Licensing standards there could therefore be enforced more rigorously than we do, and there is in fact frequent turnover among the British programmers.

More generally, cable ownership of physical facilities could be kept separate from a channel leasing function, handled by brokers who might schedule programs and interconnect systems. Facility ownership could be separated also from program production, syndication, and distribution. So that we could have at least three major components—owners, brokers, programmers—besides the manufacturers of cable hardware and aside, finally, from the commonly owned infrastructure (the ducts). In railroads, vertically integrated operating and hardware companies at one point put virtually all components of that service under joint ownership. Even today, the railroads own the rail beds, the rolling stock, the rails and the terminals. But in cable, no need.

How far did Sloan (or FCC) probe these issues? Cable owners were not entirely excluded from programming, as already seen, though they were severely limited.⁶⁴ No attempt was made to separate out brokers for scheduling and interconnection, although this could emerge subsequently. However, there *were* explicit attempts to insure unrestricted vertical access by programmers, so that the question is whether there will be enough such entities.

Stated otherwise, the separate ownership of transmission and program facilities in cable should make it easier to maintain non-discriminatory vertical access there than in broadcasting where licensees own both components jointly *and* hold rights to use the public spectrum. There are in fact numerous independent programmers in broadcasting, but this neither salvaged the UHF service nor insulated the whole nonnetwork system of advertisers, programmers and independent stations, against the major TV networks, their restrictive trade practices and long-term talent contracts.

In cable, the greater divorcement of transmission from program

⁶³ See, Levin's Reply Comments in FCC Docket No. 18110, August 1971, p. 17, and citations in footnote 18 therein.

⁶⁴ Sloan Report, pp. 147-48.

facilities and the far greater channel capacity per market should prevent any routine repetition of the above unless exclusionary pricing of leased channels goes unchecked. On that score, Sloan's (and the FCC's) prescriptions for nondiscriminatory vertical access are not enough; explicit buttressing preconditions are also essential to make such competitive threats credible.⁶⁵

As noted, two neglected preconditions are nonexclusive cable franchises and guaranteed access to commonly owned ducts and the subscriber's "black box."⁶⁶ Without such safeguards, direct rate regulation would probably be needed to insure effective vertical access. However, Sloan did adopt a modified common carrier concept, without rate regulation and without totally excluding systems owners from programming. It ostensibly did so to optimize the needed inflow of venture capital in cable's early growth stages.⁶⁷

What one really misses here is any decisive analysis of outright divorcement in cable. If we break out ownership of the assets from the operating rights, will rents diminish because of greater exertion for service? Will this happen because regulators now "mean business" and are more likely to revoke franchises than at present, or to let them lapse at renewal, and because the licensees know this? That is, does the divorcement help insure fulfillment of public interest requirements by making licensee selection and renewal more rigorous? Or will the contrary be true? If systems owners (with the transmission facilities) collect all the rents, will the programmers then simply lack the wherewithal and incentives to put on more public service? And if nonexclusive franchises and guaranteed access to the ducts discipline the price-setters, will there be any rents at all left to tax for public service? Such are the questions Sloan failed to ask, let alone to answer.

Still another issue needing more attention was whether common carrier control can stop short of content control if a rate regulation sufficiently lax to permit the inflow of needed capital will operate to price militant minority claimants out of the market when demand exceeds supply. To avoid equilibrium pricing at the expense of minority access may well require the common carrier itself to ration access, and the question is whether first come, first served will be any adequate criterion. For if cable is to achieve audience flow, essential to its value as an advertising medium, it must clearly schedule its programs lest potential users turn elsewhere. Once the system owner enters scheduling, he necessarily considers content and may well have to choose accordingly among competing claimants of channel access.

So there may well be an access problem regardless of who owns and controls. When deprived late-comers complain and demand extra capacity, some judgment on content seems unavoidable although common carrier status might still minimize potential restraints better than any alternative option.

A further wrinkle relates to Bell's possible massive role in installing a national cable grid. Sloan sought to determine Bell's in-

⁶⁵ The credibility of competitive threats is a problem of more general relevance in regulated communications. See [6], ch. 11.

⁶⁶ See note 29 *supra*.

⁶⁷ Sloan Report, pp. 147-48.

terest in and capabilities for such an investment. The thereby to generate the speediest creation of an expensive ture by an entity uniquely qualified to do so. The m regulatory stumbling block was the Justice Department' consent decree aimed to prevent Bell from exploiting le Western Electric in equipment manufacturing. The qu whether cable was itself a common carrier, or a form of br If the former, Bell could theoretically enter, whereas if she was clearly excluded.⁶⁸ On one hand, cable's abunda capacity creates common carrier characteristics, but originates its own programming.

Unless the consent decree is explicitly modified, or syste entirely excluded from programming, Bell's eligibility will certain. However, it is by no means self-evident that she v to enter under *any* conditions, or should be encouraged Sloan's spirited attempt to assess these issues deserved pi gence on the merits.⁶⁹

In sum, common carrier status has been urged to diversity and access to cable television without tempting owner to reduce competition to his own programming; or t the programming he fears might offend potential subscril opposed such status now for impeding optimal cable If owners could not control channel usage during the form: allegedly capital would not be forthcoming to lay the o so Sloan opted for a system attractive to investors today vertible into common carriage later.⁷¹ Meanwhile, div access would both presumably be insured via mandat purpose channels, leaving the future to decide about out mon carrier status.

While short of a full-fledged common carrier appr principle of mandatory cross subsidy was a big step b current regulatory doctrine in broadcasting. How succes be as a substitute for such policies as the fairness doctrine book type program service standards will depend on the ne types that emerge.

⁶⁸ See Kestenbaum, "Effect of AT&T Consent Decree on AT&T in Cable Communications" (memo to Sloan Commission), Oct. 12, 1970, and "Role of Telephone Companies in Cable Communications" (memo to Sloan Commission, Oct. 1970).

⁶⁹ It is now locked in the Minutes of the Third Meeting of the Sloan Commission, Oct. 15-16, 1970, and in the unpublished papers cited *supra*.

⁷⁰ *Sloan Report*, pp. 147-48.

⁷¹ *Ibid*, p. 148.

References

1. BARRON, J. "Access to the Press—A New First Amendment Right." *Law Review*, Vol. 80, No. 8 (June 1967), pp. 1641-1678.
2. COMANOR, W. AND MITCHELL, B. "Cable Television and the Impact on the Press." *The Bell Journal of Economics and Management Science*, Vol. 1 (Spring 1971), pp. 154-212.
3. FREDERIKSEN, H., *Community Access Video*. Berkeley: Book People Press, 1970.
4. LEVIN, H. *Broadcast Regulation and Joint Ownership of Media*. New York: New York University Press, 1960.

5. ———. "Competition, Diversity, and the TV Group Ownership Rule." *Columbia Law Review*. Vol. 70, No. 5 (May 1970), pp. 791–835.
6. ———. *The Invisible Resource: Use and Regulation of the Radio Spectrum*. Baltimore: Johns Hopkins Press, 1971.
7. PARK, R. "The Growth of Cable TV and Its Probable Impact on Over-the-Air Broadcasting." *The American Economic Review*, Vol. 61, No. 2 (May 1971), pp. 69–73.
8. POSNER, R. "The Appropriate Scope of Regulation in the Cable Television Industry." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 1 (Spring 1972), pp. 98–129.
9. ———. "Taxation by Regulation." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 22–50.

THE BELL JOURNAL OF ECONOMICS AND MANAGEMENT SCIENCE

AUTUMN '73

VOL. 4, NO. 2.

EDITOR:

Paul W. MacAvoy, *Massachusetts Institute of Technology*



ASSOCIATE EDITORS:

Dale W. Jorgenson, *Harvard University*

Harvey J. McMains, *American Telephone and Telegraph Company*

William F. Sharpe, *Stanford University*

ASSISTANT TO THE EDITOR:

Leigh E. Tripoli, *American Telephone and Telegraph Company*

TABLE OF CONTENTS

ARTICLES

- 369 Proof that properly discounted present values of assets vibrate randomly
P. A. SAMUELSON
- 375 Cartels, competition and regulation in the property-liability insurance industry
P. L. JOSKOW
- 428 The comparative statics properties of the theory of the regulated firm
D. L. McNICOL
- 454 Alternative regulatory policies for dealing with the natural gas shortage
P. W. MACAVOY AND R. S. PINDYCK
- 499 Additive nonhomogeneous production functions in telecommunications
E. F. SUDIT
- 515 Some aspects of optimal pricing for telecommunications
L. SQUIRE
- 526 Consumption externalities and imperfect corrective pricing
P. A. DIAMOND
- 539 Market changes in the computer services industry
R. S. BOWER

SYMPOSIUM ON THE NEW URBAN ECONOMICS

- 593 Notes on the new urban economics
E. S. MILLS AND J. MACKINNON
- 602 Congestion cost and the use of land for streets
R. M. SOLOW
- 619 Optimum vs. equilibrium land use pattern and congestion toll
Y. ORON, D. PINES, AND
E. SHESHINSKI
- 637 The optimum factory town
A. DIXIT

SHORT ARTICLES, COMMENTS AND REVIEWS

- 655 A further analysis of provincial trucking regulation
J. PALMER
- 665 Investment behavior in the U. S. telephone industry—1949 to 1968
U. SANKAR
- 679 Ralph Turvey's Economic Analysis and Public Enterprises
W. G. SHEPHERD
- 682 George C. Eads' The Local Service Air Line Experiment
P. VERLEGER
- 685 The FTC staff's Economic Report on Conglomerate Merger Performance
J. F. WESTON
- 690 Leonard Waverman's Natural Gas and National Policy: A Linear Programming Model of North American Natural Gas Flows
D. KENDRICK
- 693 Jan Mossin's Theory of Financial Markets
M. E. RUBINSTEIN
- 700 Omissio

Proof that properly discounted present values of assets vibrate randomly

Paul A. Samuelson

Institute Professor
Massachusetts Institute of Technology

Even the best investors seem to find it hard to do better than the comprehensive common-stock averages, or better on the average than random selection among stocks of comparable variability. Examination of historical samples of percentage changes in a stock's price show that, when these relative price changes are properly adjusted for expected dividends paid out, they are more or less indistinguishable from white noise; or, at the least, their expected percentage movements constitute a driftless random walk (or a random walk with mean drift specifiable in terms of an interest factor appropriate to the stock's variability or riskiness). The present contribution shows that such observable patterns can be deduced rigorously from a model which hypothesizes that a stock's present price is set at the expected discounted value of its future dividends, where the future dividends are supposed to be random variables generated according to any general (but known) stochastic process. This fundamental theorem follows by an easy superposition applied to the 1965 Samuelson theorem that properly anticipated futures prices fluctuate randomly—i.e., constitute a martingale sequence, or a generalized martingale with specifiable mean drift. Examples demonstrate that even when the economy is not free to wander randomly, intelligent speculation is able to whiten the spectrum of observed stock-price changes. A subset of investors might have better information or modes of analysis and get above average gains in the random-walk model; and the model's underlying probabilities could be shaped by fundamentalists' economic forces.

■ Consider a random vector sequence: $\dots, X_t, X_{t+1}, \dots, X_{t+T}, \dots$. The dividend of a particular common stock, say General Motors, might be the i th component of that vector: $\dots, x_{it}, \dots, x_{i, t+T}, \dots$; and the j th component might, as in 1965 Samuelson,¹ denote the price of spot wheat at time t . Under some known stochastic process generating the random variables, there will be defined basic conditional probabilities

$$\begin{aligned} \text{Prob}\{X_{t+T} \leq x_{t+T} | X_t = x_t, X_{t-1} = x_{t-1}, \dots\} \\ = P_T(x_{t+T}; x_t, x_{t-1}, \dots; t) \quad (1) \end{aligned}$$

Paul A. Samuelson received the B.A. from the University of Chicago in 1935 and the M.A. and the Ph.D. from Harvard University in 1936 and 1941, respectively. He is Institute Professor at the Massachusetts Institute of Technology, where his research embraces analytic micro and macro economics.

The author owes thanks to the National Science Foundation for financial aid, and to Kathryn Kaeppler for valuable assistance.

¹ See [3].

1. Review

and conditional expected values

$$\begin{aligned} E\{X_{t+T} | X_t = x_t, X_{t-1} = x_{t-1}, \dots\} &= {}_{t+T}Y_t \\ &= \int_{-\infty}^{\infty} x P_T(dx; x_t, x_{t-1}, \dots; t) \\ &= {}_{t+T}F_t(x_t, x_{t-1}, \dots) \end{aligned} \quad (2)$$

$$\begin{aligned} E\{{}_{t+T}Y_{t+1} | X_t = x_t, X_{t-1} = x_{t-1}, \dots\} \\ &= \int_{-\infty}^{\infty} {}_{t+T}F_{t+1}(x_{t+1}, x_t, \dots) P_1(dx_{t+1}; x_t, x_{t-1}, \dots; t) \\ &= E\{{}_{t+T}Y_{t+1} | {}_{t+T}Y_t\} \text{ for short.} \end{aligned} \quad (3)$$

Here a Stieltjes integral is written as $\int_{-\infty}^{\infty} f(x)g(dx)$; and when x is a vector, a *multiple* Stieltjes integral is written as $\int_{-\infty}^{\infty} f(x)g(dx)$.

The two basic 1965 theorems can now be recapitulated.

Theorem 1: For $\tau > t$, the sequence $({}_tY_t, {}_tY_{t+1}, \dots)$ has the martingale property

$$E\{{}_tY_{t+k} | {}_tY_t\} = {}_tY_t, \quad (k = 1, 2, \dots, \tau - t). \quad (4)$$

Theorem 2: For the “discounted” sequence,

$$\begin{aligned} {}_tZ_t &\equiv {}_tY_t / \prod_{j=1}^{\tau-t} \lambda_{t+j} \\ E\{{}_tZ_{t+1} | {}_tZ_t\} &= \lambda_{t+1} {}_tZ_t \\ E\{{}_tZ_{t+k} | {}_tZ_t\} &= \lambda_{t+1} \lambda_{t+2} \dots \lambda_{t+k} {}_tZ_t. \end{aligned} \quad (5)$$

2. Expected present discounted values

■ Suppose that the i th component of the vector X_t represents the dividend of a given stock that is to be paid out at time t . Then if $\lambda_{t+1} - 1$ is the interest rate paid at the end of period t on each dollar invested at time t , and if x_{it} were a nonrandom sequence, the classical Fisher present discounted-value rule of capitalization (slightly generalized) defines the value of a stock as

$$V_t = \sum_{T=1}^{\infty} (x_{it+T} / \prod_{j=1}^T \lambda_{t+j}) \quad (6)$$

$$V_{t+1} = \lambda_{t+1} V_t - x_{it+1}. \quad (7)$$

If $\lambda_t \equiv 1 + r$, the above denominator takes on the more familiar form $(1 + r)^T$.

But now revert to the supposition that x_{it+T} , and hence V_t , are random variables; and assume that *the market capitalizes the stock at the expected value of V_t , namely at v_t defined by*

$$v_t = E\{V_t | X_t = x_t, X_{t-1} = x_{t-1}, \dots\} = \sum_{T=1}^{\infty} {}_{t+T}Z_t \quad (8)$$

$$E\{v_{t+1} | v_t\} = \sum_{T=2}^{\infty} E\{{}_{t+T}Z_{t+1} | {}_{t+T}Z_t\}. \quad (9)$$

Now, by simple use of the principle of superposition, we can derive from (5) our needed generalization or corollary of Theorem 2,

namely that stock prices themselves have a martingale or random-walk property.

Theorem 3. If stocks are capitalized at their expected present discounted values defined by (8) and (9), then

$$E\{v_{t+1}|v_t\} = \lambda_{t+1}v_t - E\{x_{t+1}|X_t = x_t, X_{t-1} = x_{t-1}, \dots\}. \quad (10)$$

Clearly (10) is the fundamental stochastic generalization of the fundamental nonstochastic relation (7). Note that it holds even for the Pareto-Lévy distributions that lack a finite variance but possess a defined first moment.

Proof of the theorem follows immediately from substituting Theorem 2's relation (5) into each term of (9) and then identifying what remains by use of (8).

■ Suppose that the ratio of dividend to earnings is a constant payout fraction. Let earnings at time t be proportional to a random variable satisfying an independent multiplicative relation. Then we can deduce that dividends will be generated by the stochastic process

$$x_{t+T} = x_t Z_1 \dots Z_T, \quad (11)$$

where the Z 's are positive random variables subject to uniform and independent probability distributions

$$\text{Prob}\{Z_i \leq z\} = P(z) \quad (12)$$

$$E\{Z_i\} = \theta, \quad E\{x_{t+T}\} = \theta^T x_t$$

$$E\{\log Z_i\} = \mu < \log \theta, \quad \text{Var}\{\log Z_i\} = \sigma^2.$$

Finally, assume a constant interest rate, $\lambda_t \equiv 1 + r > \theta$, which is large enough to keep v_t a finite converging series

$$v_t = x_t \left[\frac{\theta}{1+r} + \frac{\theta^2}{(1+r)^2} + \dots \right] = x_t \theta (1+r-\theta)^{-1} \quad (13)$$

$$\text{Prob}\{v_{t+1}/v_t \leq z\} = P(z) \quad (14)$$

$$E\{v_{t+1}|v_t\} = \theta v_t \quad \text{from (12)}$$

$$= (1+r)v_t - E\{x_{t+1}\} \quad \text{from (10)}.$$

Actually this model generates the economic or multiplicative Brownian motion of Osborne and Samuelson² with the asymptotic log-normal distribution

$$\lim_{T \rightarrow \infty} \text{Prob} \left\{ a \leq \frac{\log(v_{t+T}/v_t) - \mu T}{T^{1/2} \sigma} \leq b \right\} = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}s^2} ds \quad (15)$$

and its price changes have the white-noise property

$$E\{\log v_{t+1} - \log v_t - \mu\} \equiv 0 \quad (16)$$

$$\text{covariance}\{\log v_{t+T}, \log v_t\} \equiv 0, \quad T > 0. \quad (17)$$

Granger³ has arrived at similar results, including the interesting case where variables are generated as the (possibly infinite) sum of

² See [2] and [4], respectively.

³ In [1].

3. Example of Brownian ramble

white-noise random variables. Shiller⁴ also offers valuable related contributions, particularly in connection with prediction algorithms and also the term structure of interest rates.

4. Probabilities that obey economic law

■ A second model provides an interesting contrast to the endless wandering of the above model. In it, earnings and dividends continue to have a probability distribution that stays within the same general central range; thus dividends have an ergodic distribution that is determined by economic law, by the fundamentals of the industry's resource scarcities and the capacity of its goods to meet peoples' needs and demands. But, and this is the beauty of the present martingale process, the movement of the stock price that capitalizes these determinate dividends is itself a white-noise generalized martingale!

Specifically, let dividends satisfy a damped autoregressive process

$$\log x_{t+1} = a \log x_t + \eta_t, \quad |a| < 1, \quad (18)$$

where η_t is an independently and uniformly distributed random variable, with $\text{cov}(\eta_t, \eta_{t \pm k}) = 0$ for $k \neq 0$.

Then, for $|a| < 1$,

$$\lim_{T \rightarrow \infty} \text{Prob}\{x_{t+T} \leq x | x_t = y\} = \lim_{T \rightarrow \infty} P_T(x, y) = P(x), \quad (19)$$

a limiting ergodic probability distribution that is independent of initial value for x_t and which is not log-normal.

Even though dividends and their changes have a nonwhite spectrum, with nonvanishing covariance $\{x_t, x_{t \pm k}\}$, the martingale property of Theorem 3's (10) will still be valid. Thus, if the corporation had zero dividend payments over a time interval, and the λ_{t+} discount factor were at or near unity, the spectrum of $v_{t+k} - v_t$ would be white, in the sense of zero first-order autocorrelation and zero expected values.

The present case of an ergodic probability distribution differs significantly from the log-normal models upon which so much of warrant and option valuations has been based. As applied to calls, which are typically warrants *protected* for dividend payouts, the difference is not so great. Indeed, as my colleague Robert Merton reminds me, even for the present model, once we ask what will be the cumulative value over time of a portfolio that invests back all dividends in this company's common stock, the relevant probability distribution derived from (19) will have properties much like that of (15). In fact, in the following special case, we shall have exactly the same form as (15).

Suppose the corporation selects its optimal algebraic dividend payout so as to leave within the company only that sum of wealth or money which can optimally earn more there than elsewhere. (If the indicated dividend is negative, think of the corporation as selling new shares; for that matter, transaction costs and tax complications aside, a corporation might choose always simply to buy shares *algebraically* in the open market, so that any positive dividend situation would work itself out in each of my shares' becoming more valuable.) Suppose further, for simplicity, that *ex ante* always the

same total wealth is to be left in the company: all the random events of the period just past show up in the variable algebraic dividend. Finally, let the relevant interest rates be constant, $\lambda_i \equiv 1+r$. Then each dollar left invested and reinvested in this company will be subject to the multiplicative probability distribution of (11)'s form; and (15)'s log-normal limit will apply. Even if the amount the company is to reinvest is not completely independent in probability from period to period, the white martingale property assures zero autocorrelation and unbiased means; consequently a slight generalization of the central-limit theorem, to unautocorrelated rather than independent added variates, ought still to enable derivation of a log-normal limit.

■ One person, too small to affect market prices appreciably, could make systematic speculative gains in excess of those shown in (10), if he had more or better information or a better way of evaluating existing information. This would enable him to improve upon the probability distribution of (1). Thus, suppose at time t he could know x_{t+1} exactly, or have a more accurate way of estimating it than from $P_1(x_{t+1}; x_t, x_{t-1}, \dots; t)$.

An example would be where this investor had private knowledge, or private recognition, of an additional datum m_t , in terms of which he has the probability distribution $Q_1(x_{t+1}; x_t, x_{t-1}, \dots; m_t; t)$ with the property that $P_1(x_{t+1}; x_t, x_{t-1}, \dots; t)$ is the "marginal distribution" of $Q_1(\quad)$ with m_t integrated out. Suppose

$$P_1(x_{t+1}; x_t, x_{t-1}, \dots; t) \equiv \int_{-\infty}^{\infty} Q_1(x_{t+1}; x_t, x_{t-1}, \dots; dm_t; t), \quad (20)$$

and

$$Q_1(x_{t+1}; x_t, x_{t-1}, \dots; m_t; t)P_1(x_{t+1}; x_t, x_{t-1}, \dots; t)^{-1} \\ \neq \text{a function of } m_t \text{ alone.}$$

Then knowledge of m_t gives extra predictive power of x_{t+T} and of V_{t+k} . Having such knowledge when others do not is highly profitable, since depending upon the level of m_t , the stock becomes an especially good or an especially bad buy. Of course, if this private knowledge becomes widespread, the relevant $P_1(\quad)$ will become $Q_1(\quad)$ itself, with Theorem 3 and (10) holding in terms of it, and with m_t being just one more element in the relevant x_t . In summary, the present study shows (a) there is no incompatibility in principle between the so-called random-walk model and the fundamentalists' model, and (b) there is no incompatibility in principle between behavior of stocks' prices that behave like random walk at the same time that there exist subsets of investors who can do systematically better than the average investors.

References

1. GRANGER, C. W. J. "Some Implications of the Fundamentalists' Valuation Model," May 25, 1973 paper for Berlin Workshop on International Capital Markets, September 1973.

5. Qualifications

2. OSBORNE, M. F. M. "Brownian Motion in the Stock Market." *Operations Research*, Vol. 7, No. 2 (March–April 1959), pp. 145–173. Also in P. H. Cootner, *The Random Character of Stock Market Prices*, Cambridge, M.I.T. Press, 1967, pp. 100–128.
3. SAMUELSON, P. A. "Proof That Properly Anticipated Prices Fluctuate Randomly." *Industrial Management Review*, Vol. 6, No. 2, pp. 41–49. This is reproduced as Chapter 198 in Samuelson, *Collected Scientific Papers, Volume III*, Cambridge, M.I.T. Press, 1972.
4. ———. "A Rational Theory of Warrant Pricing." *Industrial Management Review*, Vol. 6, No. 2, pp. 13–39. This is reproduced as Chapter 197 in *Collected Scientific Papers, Volume III*.
5. SHILLER, R. J. "Rational Expectations and the Structure of Interest Rates." Unpublished Ph.D. dissertation, M.I.T., 1972.

Cartels, competition and regulation in the property-liability insurance industry

Paul L. Joskow

Assistant Professor of Economics
Massachusetts Institute of Technology

This paper provides a detailed study of the structure, behavior, and performance of the property and liability insurance industry in the United States. The property insurance industry is shown to possess all of the structural characteristics normally associated with competitive markets. Despite a competitive market structure, however, the property-liability insurance industry has traditionally set prices through cartel-like rating bureaus and has been subjected to pervasive state rate regulation. The study concludes that the combination of state regulation, cartel pricing, and other legal peculiarities has resulted in the use of an inefficient sales technique, supply shortages, and overcapitalization. Free entry, however, tends to drive profits toward the cost of capital. Based on recent experience in states where the competitive market is used to determine insurance rates, the study suggests a movement away from rate regulation and cartel pricing to open competition, as a means of eliminating prevailing performance problems.

■ The property and liability insurance industry had assets of \$68 billion and premiums of \$35 billion in 1971. The products sold by this industry, in the form of contingent claims against accidental property loss and liability judgments, are purchased in one form or another by virtually all economic agents in the U. S. economy. Despite the size and importance of the property insurance industry in the U. S. economy, the literature in the area of industrial organization has all but ignored it. A leading text in the area mentions the

1. Introduction

Paul L. Joskow received the B.A. in economics from Cornell University (1968) and the M.Phil. and Ph.D. in economics from Yale University (1970 and 1972, respectively). At M.I.T. the current research of this Assistant Professor of Economics includes work in government regulation, energy demand modeling, and consumer protection.

The author became interested in the property insurance industry while working on the fire protection and prevention project of the New York City-Rand Institute during the summer of 1972. He has benefitted from discussions with members of the staff of the New York State Insurance Department, the Insurance Services Office, the New York Property Insurance Underwriting Association, Peter Diamond, Paul MacAvoy, Irving Plotkin, Peter Temin, and many others. He is grateful to the College of Insurance in New York City and the Insurance Library Association of Boston for making their library facilities available to him.

This research has been partially supported by the Ford Foundation Urban Grant to M.I.T. The views expressed here are the author's sole responsibility and do not reflect those of M.I.T. or the Ford Foundation. This paper was presented at the Dartmouth Seminar on Regulation and Public Utilities in August, 1973. Supplementary appendices to the paper are available from the author.

insurance industry in passing, but only to note that it will not be a subject of discussion in the book.¹ Scherer's relegation of the industry to the "money and banking" field may be justified in the context of past research work available, but it seems somewhat unfortunate that such an important private sector of the U. S. economy has not undergone more intensive study and analysis in the context of the structure-behavior-performance rubrick of industrial organization.

The property insurance industry is also a regulated industry. Virtually every state has an insurance commission charged with supervising the rates, financial organization, and quality of service provided by the insurance companies operating within the state. The way insurance rates are set in most states, as well as the basic structure of the regulatory process, appears to be quite different from that which has evolved in the regulated industries which are traditionally studied. A recent text dealing with the economics of regulation does little more than note the Supreme Court opinion which affirmed the power of the states to regulate insurance rates.² How the insurance industry is regulated, why it is regulated, and what the effects of regulation might be are not discussed.

For a number of reasons the time appears right for an analysis of the structure, behavior, and performance of the property and liability insurance industry operating within prevailing regulatory institutions. The recent performance of the industry has come under increasing criticism. Among other things, observers point to the shortages of fire and theft insurance in many of the nation's cities, a situation which required intervention by the federal government. Similarly there has been dissatisfaction with the availability of automobile insurance at desired coverage levels and reasonable rates through the voluntary insurance market. In addition, people have been concerned that the structure of insurance rates in many lines of insurance leads to levels of self-protection and self-insurance which are far from optimal. Finally, proposals for the implementation of no-fault auto insurance are intimately related to the structure and regulation of the insurance industry.

While unsatisfactory performance inevitably leads us to ask questions about industry structure and behavior, the presence of pervasive regulation necessitates an analysis of the regulatory environment in which this industry operates and of the possible effects of regulation on industry performance (as well as on industry structure and behavior). This appears to be a propitious time for an analysis of this regulatory process because after nearly 20 years of stability, many states are in the process of changing the way in which regulation of this industry proceeds. While some states are moving toward open competition and deregulation (New York for example), other states are moving toward adopting more traditional forms of rate-of-return regulation for the property insurance industry (New Jersey). Useful insights into the effects of regulation may be of help in aiding public policy makers in deciding which direction to move in.

This study pursues a basically traditional methodology familiar in studies of industrial organization and public utility regulation. The general lack of analyses of the property insurance industry

¹ Scherer [27], p. 2.

² Kahn [12], pp. 3 and 6.

justifies an initial attempt to lay out certain structural and institutional realities. The study therefore begins with an analysis of the structure of the property insurance industry. This section deals with such questions as market concentration, entry, economies of scale, as well as production and sales organization.

The second section examines the pricing behavior of the property insurance industry in the context of the institutional arrangements which affect pricing and the provision of property insurance. The effects of antitrust law, government regulation, and formal and informal cartel arrangements on the pricing behavior of property insurance firms are presented. Section 3 examines the performance of the property insurance industry in the context of a number of performance criteria, including supply shortages, profitability, and the efficiency of the product distribution system.

The major conclusions of the first three sections are that: (1) the property insurance industry has all of the structural characteristics of a competitive market; (2) the prior approval regulatory process employs a meaningless profitability criterion which does not necessarily "protect" consumers; (3) the insurance regulatory process has been the primary cause of supply shortages; (4) supply shortages exist at the same time that there is excess capacity; (5) the prevalent form of sales distribution system—the American Agency System—is extremely inefficient, costing consumers hundreds of millions of dollars per year; (6) available profitability studies indicate that rates of return for the industry as a whole are not excessive. While the presence of easy entry of agency firms does not make this result either surprising or particularly interesting, basic methodological flaws in these studies necessitate further research into insurance firm profitability. In addition there is evidence indicating that direct writers earn substantially higher rates of return than the industry as a whole; the underwriting behavior of direct writers is shown to be consistent with profit-maximizing oligopoly behavior of a small group of low cost firms, insulated from entry, and operating in a market where prices are kept above competitive levels by the combined actions of rating bureaus and insurance regulators.

In the final section, open competition as an alternative to all forms of direct government price regulation is suggested as a public policy goal. The property insurance industry provides a unique opportunity to do more than guess at what might happen under a regime of open competition (as opposed to regulation) because of the presence of a few instances of open or near-open competition in rates, which have actually been in operation during the period of general state regulation.

■ **Basic structural characteristics.** Insurance is a method of spreading the risk of property loss among a group of individuals. The insured trades an uncertain state of the world in which he has a small probability of a large property loss and a high probability of no loss for a certain state in which he pays a small premium for insurance against loss. By purchasing insurance, the insured is able to move from a state of uncertainty to one of certainty. Insurance is generally a "bad bet." That is to say, the premium is generally greater than the expected property loss without insurance. The difference between

2. The structure of the property and liability insurance industry

premiums and losses over time is made up of underwriting and other transactions costs and the profits of insurance firms.

The insurance market is normally separated into two broad industry groups: the life insurance industry and the property and liability insurance industry. The property and liability insurance industry, which is the focus of this study, includes fire and marine insurance, extended coverage, automobile insurance, homeowners' and commercial multiperil insurance, some types of accident and health insurance, and various other types of property and liability insurance.

Automobile insurance accounted for nearly half of the total premiums written by the property-liability insurance industry in 1971.³ This was not always the case. The dominance of automobile insurance is a post-World War II phenomenon, coinciding with increased automobile usage and the advent of compulsory auto liability insurance in many states. Historically, the structure, behavior, and regulation of this industry has its roots in fire insurance, which accounted for almost 30 percent of total property-liability premiums as late as 1943. This historical evolution is critical to an understanding of the property-liability insurance industry today, with regard to price determination and government regulation, and will be discussed extensively below.

Firms selling property and liability insurance in the U. S. are of four basic organizational forms. By far the most important type of firm is the *stock company*, owned by stockholders who have invested equity capital in the enterprise. Stock companies possessed \$43 billion in assets in 1970 or 73 percent of the total assets of the property and liability industry and wrote over \$22 billion in premiums or 68 percent of total industry premiums. *Mutual companies* are corporations which are owned by their policy holders. Instead of paying stock dividends out of profits, dividend payments are often made directly to policy holders, effectively lowering their insurance rates. In 1970 mutual companies had assets of \$14 billion or 24 percent of

³

PREMIUM DISTRIBUTION BY LINE OF INSURANCE—1971

<i>Type of Insurance</i>	<i>Percent of Total</i>
Fire	6.3
Allied Lines	2.6
Homeowners	8.6
Commercial Multiperil	5.1
Ocean Marine	1.3
Inland Marine	2.6
Group Accident and Health	2.6
All Other Accident and Health	1.6
Workmen's Compensation	10.8
Miscellaneous Liability	6.9
Auto Liability	30.9
Auto Physical Damage	16.8
Aircraft	0.6
Fidelity	0.5
Surety	1.2
Glass	0.1
Burglary	0.4
Boiler and Machinery	0.4
Credit	0.1
Miscellaneous	0.6

Source: *Best's Review* [6], July 1972.

total industry assets and had premiums of almost \$9 billion or 27 percent of total industry premiums. A *reciprocal exchange* is a co-operative organization formed to share specified risks of the members of the exchange. Reciprocal exchanges had assets of almost \$2 billion in 1970 or 3 percent of total industry assets and wrote about \$1.5 billion in premiums or 4 percent of total industry premiums. About 80 percent of the premiums of reciprocal exchanges are derived from auto insurance, and the reciprocals accounted for 7.6 percent of auto premiums in 1969. Finally, *Lloyd's Organizations* are made up of groups of underwriters, each taking on a portion of the insured's risks. Lloyd's Organizations accounted for only 0.1 percent of total industry assets and 0.1 percent of total industry premiums in 1970.⁴

Property insurance is marketed in two basic ways in the United States. The method employed by most stock companies is the American Agency System. Under the American Agency System independent retailers or agents represent a number of insurers and sell insurance for these companies to the public. For their efforts, the agents receive a commission, usually a fixed percentage of the premiums written. The system of "direct writing" arose in competition with the agency system, primarily (apparently) because of the high sales costs that had resulted. Insurance companies that became direct writers sold insurance directly, either through their own salesman or through the mail. Direct writing is a very important part of automobile insurance sales, accounting for nearly 50 percent of total auto premiums in 1971, but much less important in other insurance lines. Although the importance of direct writing has been growing over time, it has not grown so fast as one might expect.⁵ An analysis of why this apparently superior marketing system has not completely taken over, and estimates of the economic losses attributable to the American Agency System, are presented in a section on industry performance below.

□ **Number of firms, concentration and entry.** *Best's* lists 840 stock companies, 311 mutual companies, 44 reciprocal exchanges, and 11 Lloyd's Organizations for a total of 1206 property-liability insurance companies operating in the United States in 1971. Most of these companies operate in more than one state and a substantial number of these companies operate nationwide. In 1971 about 80 percent of property-liability premiums were written by national agency companies or direct writers. Of 829 auto insurers analyzed for 1967, 650 of them were regional, multiregional or national and only 179 licensed to operate in only one state.⁶

⁴ The previous data on market shares of the various types of "producers" comes from *Best's Aggregates and Averages* [4], 1971.

⁵

PERCENTAGE OF TOTAL PREMIUMS WRITTEN BY
AGENCY AND DIRECT WRITING COMPANIES

	1967	1968	1969	1970	1971
National and Regional Agency Companies	71.5	70.5	69.5	69.0	68.5
Direct Writers	28.5	29.5	30.5	31.0	31.5

Source: *Best's Review* [6], July 1972.

⁶ Federal Trade Commission Report [32], p. 13.

Any insurance company must be licensed in each of the states in which it operates. The various states have different laws regarding the requirements for both incorporation in the state and for state corporation operating within the state. In general, the laws will specify a minimum amount of paid-in-capital needed to commence operations. In addition, the laws will indicate the types of securities the minimum capital is to be kept in, as well as eligible investments for other financial reserves. In New York, for example, the insurance law requires that a stock company selling fire insurance must have a minimum of \$250,000 of paid-in-capital and an additional \$250,000 of initial surplus to be organized as a fire insurance company. In addition to selling fire insurance, such a company may also sell miscellaneous property insurance (essentially extended fire insurance), water damage insurance, collision insurance, motor vehicle insurance, aircraft insurance (except liability for personal injuries), and various types of marine insurance.⁷ Out-of-state companies must meet the same minimum capital requirements to be licensed in New York, while alien companies require 200 percent of the minimum requirements of domestic corporations. Companies wishing to sell various casualty lines not included under this classification must have up similar amounts of capital and surplus to be licensed to sell those lines. In most states mutual companies have surplus requirements which are similar or identical to the total capital and surplus requirements of stock companies.

The presence of a large number of firms selling essential insurance products does not itself assure a competitive market in the industry. If a small number of firms control most of the industry, the presence of many firms may only give the illusion of competition. As a result, we would like to examine the level of concentration in the property-liability insurance industry as a whole as well as within individual lines of insurance. The concentration ratios shown below indicate the share of total property-liability insurance premiums written by the largest firm, the four largest firms, the ten largest firms, and the twenty largest firms for the nation as a whole. The definition of firm chosen here is that of the insurance company. Often individual insurance companies are part of the same parent-owned management group. Although individual companies within a group have considerable autonomy, joint directorships and other relationships make it appropriate to consider an insurance group as a single firm when examining questions of market control. Independent companies are considered as if they were groups composed of one company.

A sales measure of concentration has been chosen—sales written—instead of an asset measure so that national concentration ratios for the industry as a whole can be compared with concentration ratios for particular lines of insurance. Since asset accounts are not segregated by line, sales concentration ratios are the only ones feasible. In addition, since group insurance is being used, asset concentration ratios would probably be biased upward, reflecting the assets of particular companies in some lines which do very little actual property-liability insurance business.

The use of national concentration figures must also be

In general, purchasers of insurance are limited to obtaining coverage from companies licensed within a particular state. However, since entry into any particular state by an established insurance company is very easy, sellers can easily move into any state if profitability conditions warrant it. In addition, the large national agency firms and the direct writers account for 81 percent of property-liability premiums written nationwide and operate in almost every state. Although there will be some local variation, the property-liability insurance firms are essentially operating in a national market.

The concentration ratios presented in Table 1⁸ indicate that although there were over 1200 firms selling property and liability

TABLE 1
CONCENTRATION IN THE PROPERTY AND LIABILITY INSURANCE
INDUSTRY FOR SELECTED YEARS, % OF TOTAL NET PREMIUMS WRITTEN

YEAR	TOP GROUP	TOP 4 GROUPS	TOP 8 GROUPS	TOP 20 GROUPS
1962	4 4	16 5	29 0	48.1
1971	6 2	19.8	32.8	54.2

insurance in 1971, the top 20 groups controlled over half of the market. Comparisons with concentration ratios for other industries indicate that those for the property and liability insurance industry are relatively quite low.⁹

Table 1 does indicate, however, that the levels of concentration have been increasing over time. The rate of growth is relatively small, and even if the current growth rate in concentration continued, it would take many years before the four largest firms attained as much

⁸ Calculated from data obtained from *Best's Aggregates and Averages* [4]. Comparable group data only available since 1962.

⁹

1963 CONCENTRATION RATIOS FOR
REPRESENTATIVE INDUSTRIES
(4-Digit Industries)

<i>Industry</i>	<i>4-Firm Ratio</i>	<i>8-Firm Ratio</i>
Passenger Cars (5-digit)	99	100
Primary Aluminum	96	100
Cigarettes	80	100
Tires and Tubes	70	89
Motors and Generators	50	59
Beer and Malt Liquors	34	52
Cement	29	49
Fluid Milk	23	30
Men's and Boys' Suits	14	23
Bottled and Canned Soft Drinks	12	17

Source: U. S. Senate, Committee on the Judiciary, Subcommittee on Antitrust and Monopoly, Report, *Concentration Ratios in American Manufacturing Industry: 1963*, Part I, 1966.

The comparison of sales concentration ratios for insurance firms with those for manufacturing enterprises probably overstates the relative degree of control of the firms in the insurance industry. Insurance firms can easily expand output in the short run with little or no increase in "capacity" and both output and capacity in the long run by obtaining more equity capital if such expansion is desirable to compete with a particular company which gets out of line in terms of price. Particular firms can vary output and "capacity" much more quickly and with less cost than could a steel or automobile firm.

as 50 percent of the market. More importantly, it will be argued below that the increasing levels of concentration are indicative of a secular movement of business away from high cost firms employing an inefficient marketing technique toward lower cost, more efficient firms. Increasing concentration in this case is the result of increasingly effective competition in a market constrained by a peculiar array of regulatory and other legal constraints. As shall be argued below, the slowly increasing level of concentration is indicative of an improvement, and not an erosion of consumer welfare.

Although these aggregate concentration ratios give us some indication of overall market control by leading firms in a national market, it is worthwhile to examine market concentration for individual lines of insurance. Automobile insurance appears to have concentration ratios somewhat higher than the industry as a whole, but these are still relatively low¹⁰ and have been increasing at about the same rate as the industry as a whole. It will be argued below that these increased concentration ratios are indicative of movements of customers away from high cost firms operating under the American Agency System toward lower cost direct writers. Fire insurance possesses concentration levels almost identical to the property-liability industry as a whole (in 1971), and these concentration levels have been almost constant over time. This is indicative of the fact that fire insurance has been of declining importance to many property-liability firms and has been of only minor interest to direct writers using mass marketing techniques.¹¹ All things considered, the property-liability industry nationally and by line appears to possess an atomistic market structure.

□ **Selling costs and scale economies.** Besides paying out money for losses and loss adjustment expenses, property-liability insurance companies also incur substantial sales or underwriting expenses. In 1970, selling expenses amounted to 36.0 percent of premiums written for stock companies, with commissions alone accounting for 21.4 percent of premiums written. Underwriting expenses were as high as 44.7 percent of premiums written in 1960. These figures may have important implications for theoretical work regarding economic

10

CONCENTRATION RATIOS FOR AUTOMOBILE INSURANCE

(% of Earned Premiums by Top Groups)

<u>Year</u>	<u>Top Group</u>	<u>Top 4 Groups</u>	<u>Top 8 Groups</u>	<u>Top 20 Groups</u>
1954	4.8	17.7	27.9	45.3
1962	7.5	21.0	31.1	46.6
1971	11.3	27.3	38.8	56.6

Calculated from data obtained from *The National Underwriter*, May 5, 1972, June 7, 1963, and May 19, 1955.

11

CONCENTRATION RATIOS FOR FIRE INSURANCE

(% of Net Premiums Written by Top Groups)

<u>Year</u>	<u>Top Group</u>	<u>Top 4 Groups</u>	<u>Top 8 Groups</u>	<u>Top 20 Groups</u>
1954	6.40	17.87	29.42	49.04
1962	5.28	19.42	31.64	53.88
1971	5.06	19.37	32.98	56.73

Calculated from data reported in *Best's Aggregates and Averages* [4].

decision making under uncertainty. The assumption of perfect "no-load" insurance markets is extremely questionable. With transactions costs for property-liability insurance ranging between 35 and 50 percent of premiums written, there may be many risks which are completely uninsurable under present market conditions. Assuming that selling costs plus a risk adjusted rate of return on investment are equal to 50 percent of the premiums, risks with probabilities of loss greater than $\frac{1}{2}$ may be uninsurable. Assuming that in a competitive market the premium for a risk is equal to the expected loss plus selling expenses plus a risk adjusted profit (profits will be discussed further below), with selling expenses equal to 50 percent of premiums, risks with a loss probability greater than $\frac{1}{2}$ will only be insured at a premium greater than the value of the property. Such a property is essentially uninsurable.

The property-liability insurance industry has traditionally been criticized for having excessive underwriting costs. Innovations or improved efficiency which could lower these costs would both increase the availability of property insurance and decrease the rates on those properties which are currently insured. Stock companies have traditionally borne the brunt of the criticism regarding selling costs, because their costs have traditionally been so much higher than those of mutual companies. This difference was especially pronounced ten years ago, but the difference between stock and mutual company expense ratios has declined secularly over the past ten years.¹²

The reduction of the proportion of the premiums going to selling and other underwriting expenses for stock companies appears to be in response to a fairly clear set of market stimuli. Relatively poor profit performance after the mid 1950s (the initial years of prior approval state regulation), resulting from rapid entry into the industry of all types of firms (especially the dramatic growth of direct writers selling at reduced prices in the profitable auto lines), led many stock companies operating under the American Agency System to cut costs by increasing internal efficiency and to force lower commission rates on the independent agents. Mutual companies have been able to keep their expenses low by paying lower commissions, concentrating on a few classes of business, and having exclusive agents who sell insurance for only one company. Even in 1970, brokerage and commission expenses for stock companies were substantially higher than similar expenses for mutual companies.

12

STOCK AND MUTUAL COMPANY EXPENSE RATIOS 1960-1970
(Expenses/Premiums Written)

<u>Year</u>	<u>Stock</u>	<u>Mutual</u>
1970	36.0	33.7
1969	37.6	34.4
1968	38.9	35.1
1967	39.7	34.3
1966	40.8	34.5
1965	42.1	35.2
1964	43.6	35.4
1963	44.3	36.2
1962	44.5	34.0
1961	45.0	34.9
1960	44.7	35.2

Source: *Best's Aggregates and Averages* [4], 1971.

It is argued below that the American Agency System is an extremely inefficient sales technique compared to the alternatives of direct writing and the use of exclusive agents. It has led to substantially higher insurance costs than would be yielded by the least cost marketing technique. It has been preserved as a combined result of price regulation by state commissions, price making in concert, and a quirk in the insurance law. The law vests property rights to policyholders in the independent agent who writes the insurance. It bars the company from independently writing insurance for a customer who was originally obtained through a particular agent. This issue is more conveniently analyzed after a discussion of economies of scale in the marketing of insurance is presented. We turn to this question now.

There is a tremendous range in the sizes (in terms of premiums written as well as assets) of property and liability insurance firms in the United States. There is a substantial number of firms with total annual premiums of less than \$1 million and a few with total annual premiums of over \$1 billion. Given this large variation in firm sizes, it is of interest to ask whether or not the larger firms have a cost advantage over the smaller firms. Such information will have important implications for a discussion of barriers to entry into the property-liability insurance industry, and the evaluation of the performance of the industry, especially with regard to the impact of state price regulation and cartel-like rating organizations. Substantial cost advantages for very large scale operations could indicate substantial barriers to entry and the possibility that large firms could set prices substantially above marginal cost without provoking competitive entry. At the same time the presence of a large number of small, high cost fringe firms may indicate that the large companies, acting through the cartel-like rating organizations, have succeeded in keeping prices above the competitive level, thus protecting inefficient producers. This latter phenomenon will most likely occur only if entry at optimum size is difficult.

A rather crude analysis of possible scale economies in the "production" of property-liability insurance has been performed by Hensley.¹³ The study lacks any statistical tests, lumps American Agency companies together with direct writers, and does not correct for the "mix" of insurance lines in particular companies. His conclusion that there are moderate economies of scale may be especially sensitive to the nondifferentiation between direct writers and agency companies. Casual observation indicates that direct writers employ a more efficient sales technique and have a disproportionately high representation among large insurers. The study presented here should be an improvement on all of these counts. It offers a framework for answering the question of whether direct writing is the lower cost production technique. In addition, it provides useful empirical information for estimating the resource loss associated with the American Agency System.

The economies of scale analysis performed here relates expense ratios to measures of firm size, business character, and marketing technique for three different samples of companies: stock companies which are primarily auto insurance producers; mutual companies

which are primarily auto insurance producers; and stock companies which are primarily fire insurance producers. Each company sample is analyzed separately (the stock auto and stock fire samples for two different years), and then as a pooled regression.¹⁴ The basic form of the model used in the economies of scale analysis is the following:

$$E = F(S, M, \epsilon),$$

where

E = Expense ratio,

S = Variables measuring size of the company,

M = Variables measuring marketing characteristics, and

ϵ = Random disturbance term.

If there are economies of scale present, we would expect the expense ratio to decline with firm size. Since Hensley's analysis considered only the relationship between expense ratios and firm size, ignoring other characteristics of the firms, an initial set of regressions was run to see whether there is any "evidence" of scale economies when other company characteristics are omitted. Simple linear regressions relating the expense ratio to premium volume for three sets of data were calculated. The results for the regression equations in which premium volume enters linearly are reported below.¹⁵

Fire and Allied Lines (1971)

$$E = 34.28 - 0.5 \text{ DPREM} \quad \text{Observations: 25}$$

$$(21.22) \quad (-0.28)$$

$$R^2 = 0.003$$

Mutual Auto Insurers (1971)

$$E = 26.41 - 0.60 \text{ DPREM} \quad \text{Observations: 34}$$

$$(23.30) \quad (-1.85)$$

$$R^2 = 0.09$$

¹⁴ All of the data employed in the expense analysis came from *Best's Aggregates and Averages* [4], 1972 and 1971 editions. The fire insurance company sample consists of 26 stock companies chosen from those listed in *Best's* as predominantly fire insurance writers in 1971. Additional data were collected for the 23 of these companies for which data were also available for 1970. The mutual company sample consists of 32 companies from those listed in *Best's* as predominantly auto (participating and deviating) for the year 1971 plus three additional companies listed under different headings but writing a substantial portion of their business in auto insurance. Finally the auto stock company sample originally consisted of 39 stock companies from those listed by *Best's* as predominantly auto (participating and deviating) for 1971—data inconsistencies and omissions reduced this sample to 36 companies for 1971. Additional data were collected for 37 of the original 39 companies for 1970.

Determining which companies are direct writers was not an easy task. No generally available guide provides a concise listing. Attempts to get a listing of direct writers from *Best's*, etc., were only partially successful. As a result, it was necessary to supplement readily available information by going through corporate underwriting descriptions in *Best's Property and Liability Insurance Reports* [5], 1972 edition, company by company for each sample. Where there was some ambiguity the author's best judgment was used in making the allocations.

¹⁵ These relationships were also estimated allowing premium volume to enter quadratically.

Stock Auto Insurers (1971)

$$E = 25.58 - 0.35 \text{ DPREM} \quad \text{Observations: 35}$$
$$(19.66) \quad (-0.92)$$
$$R^2 = 0.03$$

Stock Auto Insurers (1970)

$$E = 25.92 - 0.48 \text{ DPREM} \quad \text{Observations: 37}$$
$$(22.55) \quad (-1.13)$$
$$R^2 = 0.04,$$

where

E = Expense ratio

DPREM = Direct premiums written.

Even this naive formulation of the relationship between expenses and firm characteristics yields only weak evidence of scale economies. Although the coefficient of the premium volume variable was negative in all cases examined, it was only close to being significant at the 5-percent level in the mutual auto insurance sample. Allowing premium volume to enter quadratically did not yield any better results, nor did the use of net premiums instead of direct premiums. Even if no important explanatory relationships had been left out of these equations, the notion that there are economies of scale in the "production" of insurance is given only weak to moderate support by these results.

In addition to premium volume there are at least two prospective characteristics of the insurance firm which can be reasonably expected to affect expenses. One involves the firm's reinsurance activities. When a firm takes on insurance from other firms (reinsurance) it must make additional commission payments. When it cedes reinsurance to other firms, it receives commission income.¹⁶ Therefore we would expect, *ceteris paribus*, that firms whose net premiums (after all reinsurance activity) are large relative to their direct premium writings will have higher expenses as a proportion of premiums written than firms whose net premiums are low relative to their direct premiums written.

A second important consideration is the method of sales. Some of the companies in these samples are direct writers and are in essence employing a different production activity than the agency firms. Although we would expect the least cost technique to be adopted in a competitive market in the long run, and would therefore expect the existence of two different techniques in competitive equilibrium to imply trivial cost differences, the behavior of this industry in this regard may be far from competitive. In particular, it is expected that direct writers will exhibit significant cost advantages, and that the agency firms have been able to survive because of a number of market imperfections. Finally, it is possible that direct writers or agency companies, but not both, exhibit economies of scale.

As a result, a far more reasonable production cost model is the following:

$$E = a + b \text{ DPREM} + c \text{ INTER} + d \text{ RATIO} + e \text{ DWRITE},$$

where

E = Expenses as a percentage of premiums written or earned,

$DPREM$ = Direct premium volume,

$INTER$ = Direct premium volume of direct writers and zero for agency companies,

$RATIO$ = Net premiums/direct premiums, and

$DWRITE$ = Dummy variable = 1 direct writers
= 0 otherwise.

The *a priori* expectations for the coefficients are the following.

- b*: This coefficient will be zero if economies of scale are absent and less than zero if economies of scale are present for agency companies.
- c*: This coefficient will be negative if direct writers have "more" scale economies than agency firms and positive if they have less.
- d*: This coefficient should be positive, indicating the premiums net "reinsurers" must pay.
- e*: This coefficient should be negative, indicating the cost savings associated with direct writing.

*Stock Auto (1970)*¹⁷

$$E = 21.41 - 0.28 DPREM + 0.34 INTER + 8.10 RATIO - 11.51 DWRITE$$

(7.35) (-0.11) (0.14) (2.69) (-6.33)

$$R^2 = 0.61$$

Stock Auto (1971)

$$E = 20.70 + 0.86 DPREM - 0.78 INTER + 7.49 RATIO - 10.68 DWRITE$$

(6.66) (0.34) (-0.30) (2.72) (-4.85)

$$R^2 = 0.54$$

Stock Auto (1970 + 1971)

$$E = 21.38 + 0.33 DPREM - 0.24 INTER + 7.46 RATIO - 11.08 DWRITE$$

(10.60) (0.19) (-0.14) (3.88) (-8.07)

$$R^2 = 0.57$$

Mutual Auto (1971)

$$E = 14.63 - 0.62 DPREM + 0.62 INTER + 16.36 RATIO - 14.08 DWRITE$$

(2.45) (-1.99) (1.99) (2.49) (-5.46)

$$R^2 = 0.55$$

¹⁷ T-statistics appear in parentheses under the estimated coefficients.

Stock Fire (1971) (no Direct Writers in sample)

$$E = 33.74 - 0.45 \text{ DPREM} + 0.38 \text{ RATIO}$$

(16.70) (-0.25) (0.45)

$$R^2 = 0.01.$$

The results of these regression estimates are of great interest. Evidence of economies of scale among stock companies and among direct writers is completely absent. There is evidence that mutual auto insurers exhibit scale economies,¹⁸ although the direct writers in that sample do not. The conclusion must be that stock property and liability insurance companies and all types of direct writers, the types of companies which sell the vast majority of property and liability insurance, exhibit no significant scale economies. In addition, the *RATIO* variable and the dummy variable denoting direct writers always have the expected signs and are significant at the 5-percent level for each sample except that for the stock fire insurance companies.

This analysis leads to the conclusion that the production of property insurance is characterized by constant returns to scale. Large companies do not appear to be able to produce insurance less expensively than small companies. We observe higher costs for firms which assume reinsurance from other firms, but this can be viewed as a secondary production activity of insurance firms. Finally, direct writers appear to exhibit significant and substantial cost savings over agency companies.¹⁹ This phenomenon will be discussed more fully in a section below.

□ **Entry.** The difficulty of entry into an industry has important implications for the ability of the market to operate close to the competitive norm, especially in industries with high concentration ratios.²⁰ If a group of insurance companies act in concert through a cartel, attempting to hold prices well above competitive levels in the presence of only modest entry barriers, such efforts will ultimately be defeated by the entry of new firms. If existing firms wish to deter new entry, they must hold the margin between price and cost to a level less than the cost advantage existing firms have over potential entrants. Large deviations of price above the competitive level will tend to result only if the industry has a small enough number of firms to fix prices covertly or legal mechanisms which can enforce cooperative pricing *and* high barriers to the entry of new firms. Since concentration ratios are low in the property-liability industry, cartel-like pricing which might exist would have to rely on legal mechanisms supporting such pricing behavior. We will turn to the question of cooperative pricing

¹⁸ This result may have emerged because the larger companies in the mutual company sample are all direct writers.

¹⁹ This evidence should not be construed as implying that small companies are not *riskier* than large companies. The nature of insurance makes it essential that companies be large enough to exploit the law of large numbers. The opportunities for reinsurance facilitate this convergence greatly. An analysis of the loss experience of insurance firms and how it relates to company characteristics like size is now under way. In the context of the discussion presented here, the question of whether existing small firms exhibit significantly more loss variability than large firms will be analyzed.

²⁰ Bain [3].

in the next section, but it is shown below that even if regulation and rating bureaus can facilitate cooperative pricing, entry has been extensive and entry barriers are low to moderate, making it extremely unlikely that a substantial margin between prices and costs, resulting in excess profits, could persist.

The record for the past twelve years indicates that there has been continuous and substantial entry of corporations into the property-liability insurance industry. For the period 1960 to 1971 a total of 336 new companies are reported as entrants into the property-liability industry. The rate of entry has varied from a low of 14 companies in 1966 to a high of 51 companies in 1961.²¹ This substantial amount of entry appears to be the result of rapid growth in the demand for property and liability insurance (especially auto insurance), a regulatory system which effectively pegs prices at a level that is on average above marginal cost,²² and very low entry barriers. We examine entry barriers next and turn to the regulatory process in the next section.

Economies of scale have been examined in the previous section. Since they are nonexistent (at least for agency companies), they do not appear to be an important barrier to entry. In many American industries, even where substantial economies of scale in production do not exist, potential entrants face the problem of getting their products recognized by consumers in a market characterized by heavy product differentiation and entrenched brand loyalty. This has not been too much of a problem for potential entrants into insurance in the past for a number of reasons. The insurance product itself—the policy—is essentially identical from company to company within a state since most policy forms (except for special risks) are mandated by state law. Under the American Agency System consumers do not shop around for a company they recognize or like, but rather seek an agent who will try to get the customer insurance coverage from one of the companies that he represents, often at a price fixed by a rating bureau and adhered to by a large proportion of the companies in the market. Under the American Agency System the consumer must rely on his agent to find a company and secure the best price if there happens to be some price variation among com-

21

ENTRY INTO THE PROPERTY AND LIABILITY INSURANCE
INDUSTRY IN THE U. S. 1960-1971

<u>Year</u>	<u>Number of New Companies</u>
1960	27
1961	52
1962	41
1963	35
1964	23
1965	18
1966	14
1967	24
1968	21
1969	15
1970	32
1971	34
Total	336

Source: *Best's Insurance News*, 1961-1971.

²² See pp. 394-395, below.

panies. The company writing insurance under the American Agency System is therefore faced with minimal "product recognition" costs and instead must contact independent agents who will be willing to market the company's policies. The costs of "plugging in" to this existing marketing network should be very low.

Company identification is probably much more important for companies wishing to enter the market as direct writers. New companies must make expenditures to make consumers aware of their products. Mass advertising and direct mail campaigns are favorite devices for the establishment of product identification and for contacting potential customers. It is difficult to say whether this is an important barrier or not. The available information is quite ambiguous. For example, GEICO was able to enter the market at a relatively small size, has had extremely good earnings performance, and has become one of the largest automobile insurers in the country. However, the high levels of concentration among direct writers, as well as apparently consistent profit performance above the industry average, indicate that barriers to entry for direct writers may be quite high.²³ Substantial initial funding may be necessary to get a direct writer set up in the market.²⁴

Artificial constraints on entry are also quite small. Rating bureaus must now be open to all to subscribe to or join. State licensing appears to be easily available to any company that can meet the minimum capital requirements and other statutory restrictions. Some states, however, still impose special taxes on companies domiciled in other states which would raise the costs of a foreign corporation. This is an effort by some states to protect domestic corporations and may deter entry by foreign or alien firms.

Entrants seeking to enter markets at deviated or bureau rates may experience opposition from rating bureaus which have often been recognized as aggrieved parties.²⁵ In addition, associations of insurance agents may attempt to restrict the entry of mutual companies, participating stock companies, and direct writers.²⁶

All things considered, most of the traditional barriers to entry discussed so far appear to be quite low. The one remaining barrier is the capital requirement for entering the insurance market. We turn to this question now.

The capital requirements of a company seeking to become incorporated or to enter a state where it has not been licensed previously are not very high. State insurance laws normally set minimum amounts of paid-in-capital and surplus as a requirement for obtaining a license to sell insurance. The capital requirement varies

²³ These issues are discussed more fully below. The suggestion that there are moderate to high barriers to entry does not necessarily contradict the earlier finding that there are not economies of scale exhibited by direct writers. Those firms successfully operating in the market and included in our samples have probably been operating long enough to have achieved minimum efficient size. The entry barriers themselves are more of the nature of "threshold" barriers requiring substantial expenditures to gain even minimal market recognition.

²⁴ Allstate was able to use the existing Sears sales network, State Farm was able to use farm bureau contacts, and GEICO concentrated on a restricted group of customers at first.

²⁵ U. S. Senate, Subcommittee on Antitrust and Monopoly [31], Part 2, pp. 925, 930-931, 940.

²⁶ *Ibid.*, pp. 920-922.

from state to state and depends on the number of lines of insurance the company wishes to write. In New York State a stock company must have paid-in-capital and surplus of over \$3,000,000 to write all nonlife lines. To write auto insurance it would require paid-in-capital and surplus of \$500,000, but this would also entitle the company to write several other lines. Similar rules apply to mutuals.

All things considered, barriers to entry, at least for agency companies, appear to be low. Direct writers apparently are faced with higher barriers to entry, attributable primarily to advertising and other “product recognition” expenditures and the recruitment of a core of sales personnel. We would therefore expect that as long as the rating bureaus can keep insurance rates above the costs of agency companies, free entry will drive rates of return on capital toward the competitive level. However, we shall see below that such a “normal” profit equilibrium is characterized by prices above the competitive level, excess capacity in the industry as a whole, and selective underwriting (and above normal profits) by the low cost direct writers.

□ **Market structure: conclusions.** The property-liability insurance industry possesses the structural characteristics normally associated with the idealized competitive market: a large number of firms, operating in a market with low concentration levels, selling essentially identical products, provided at constant unit costs and with ease of entry of new and potential competitors. The only deviation from the ideal is presented by the possibility that direct writers may face high entry barriers. It is indeed difficult to find too many other industries which conform more closely to the economist’s idealized competitive market structure.

When we turn from market structure considerations to the behavior of the property-liability insurance industry, we do not find a continuation of the competitive ideal, however. Insurance rates have traditionally been set in concert through rating bureaus, with the rates subject to some form of state regulation in most cases. To understand the pricing behavior of the property-liability industry today, and to evaluate proposed changes in state regulation, the fault system and other public policy issues, we must first examine the development of rate making in concert, antitrust, and regulatory policy as it pertains to this industry.

■ **The development of government policy toward the insurance industry.** Although a majority of the property and liability insurance sold in the United States today consists of auto and homeowners’ multiperil insurance, industry price setting behavior and government regulatory policy have their origins in fire insurance. Since fire and allied lines insurance was a very important component of the property insurance industry as recently as World War II, this is not too surprising. To understand how current pricing behavior and regulatory control evolved, we must take an excursion back into history.²⁷

The history of fire insurance rating and regulation is a direct con-

3. Pricing behavior in the property and liability insurance industry

²⁷ A more detailed discussion of the history of government policy toward the insurance industry is contained in my paper, “Regulation and Deregulation in the Property Insurance Industry, 1870–1973,” unpublished.

sequence of the essentially noncompetitive fire insurance market, dominated by cartels and essentially exempt from the federal anti-trust laws, that existed through most of the first half of this century.

The development of cooperative fire rate making goes back to the beginning of the 19th century:

It began with local boards, one of which was organized in New York in 1819. In 1866, the National Board of Fire Underwriters was organized to establish and maintain uniform premium rates and to control agents' commissions During profitable periods in the fire insurance business, however, companies violated their membership agreements by ratecutting.²⁸

The National Board was discontinued in 1877 as a countrywide rating organization and replaced with local and regional rating organizations. The National Board was essentially a cartel meant to fix rates, to overcome the "bad effects" of competition—namely insolvencies—and generally to "stabilize" the market. It had difficulty succeeding as a cartel because membership was voluntary and it was difficult to force a large number of firms to adhere to the bureau rates.

The development of the "compact system" succeeded the demise of the National Board.²⁹ Under this system local insurance agents agreed to respect uniform rates set by compact managers. The "compact system," however, was evolving at the peak of antitrust fever in the United States. From 1885 to 1907 about 20 states enacted antcompact laws to combat price fixing in fire insurance. The compact laws do not appear to have been very successful in stemming the *de facto* fixing of rates in concert.³⁰

By the late 19th Century proposals for regulation of insurance rates were made in several states. A Joint Legislature Committee, known as the Merritt Committee, was formed in New York in 1910 to study problems of fire insurance rate making. Based primarily on the findings of the Merritt Committee, an insurance law was passed in New York State in 1911 which permitted "action in concert in the fixing of fire insurance rates, but required rating associations or bureaus to file such rates with the Superintendent of Insurance."³¹ The legislation required that rates set not be unfairly discriminatory. Although the Superintendent of Insurance could evaluate rates after they were filed, prior approval was not required at this time.

There are two interesting features of government policy toward the insurance industry at this time. The primary concern of both the insurance companies and their regulators was to guard against rates that were *too low*. Competition was viewed by industry and its regulators as leading to instability and insolvencies among fire insurance firms. The regulatory agencies apparently did not view their jobs as guarding against monopolistic pricing resulting from rate making in concert, but rather as making sure that firms did not charge

²⁸ New York State Insurance Department [21], p. 69.

²⁹ The National Board continued to exist, but not as a formal rating bureau.

³⁰ Merritt Committee Report [22], reprinted in [31], p. 2791.

³¹ New York State Insurance Department [21], p. 69. The *raison d'être* of the Merritt Committee report seems to have been to convince state legislators that unregulated competition in insurance would be "destructive" with regard to the financial viability of insurance companies as well as incentives to provide rate structures which encourage fire prevention.

off-bureau rates that were too low. Competition in fire insurance rate making was viewed as being destructive and rating bureaus and regulatory agencies made sure that price competition became virtually nonexistent. No study seems to exist which shows that competition in fire insurance is any more "destructive" than in any other industry, and it appears that the evolution of regulation of the fire insurance industry stems more from an effort to protect existing firms than the interests of consumers.

The second interesting feature of public policy toward the insurance industry involves the federal antitrust statutes enacted between 1890 and 1914. The coordinated rate-setting activities engaged in by insurance firms through the rating organizations would appear to be in clear violation of the federal antitrust laws. However, the decision of the Supreme Court in *Paul v. Virginia*,³² upholding a state law subjecting an out-of-state insurance company and its local agents to a licensing requirement, contained language which was interpreted for the next 75 years as meaning that insurance could not be classified as interstate commerce.³³ As a result, the insurance industry was considered to be exempt from the federal antitrust laws.

Before 1944 the position of the fire insurance industry vis-à-vis government control was characterized by a general exemption from federal antitrust laws and direct state regulation of fire insurance rates in about 35 states. On June 5, 1944, a dramatic shift occurred in the court's interpretation of the commerce clause and implicitly of the applicability of the antitrust laws to the insurance industry tool. The court's opinion in *United States v. South-Eastern Underwriters Association*³⁴ ruled that insurance was commerce and by implication that combinations of insurance companies designed to fix rates would be in violation of the Sherman Act:³⁵

Overnight, the entire legal basis for the immunity of combinations in rate-making, the cornerstone of the fire insurance business—and hence, at that time, of the dominant segment of the property-liability insurance business—was eliminated. Moreover, doubt was cast on the system of state regulation and taxation of the insurance business. The decision precipitated widespread controversy and dismay. Chaos was freely predicted.³⁶

Undoubtedly under tremendous pressure from the insurance industry, the U. S. Congress stepped into the picture. The McCarran-Ferguson Act,³⁷ signed by President Roosevelt on March 9, 1945, declared the continued regulation and taxation of the insurance industry to be in the public interest and that the federal antitrust laws

³² 8 Wall 168 (1869).

³³ Among those cases that relied on this interpretation were *Hooper v. California* (155 U. S. 648), *Liverpool Insurance Company v. Massachusetts* (10 Wall 566), *Philadelphia Fire Insurance Association v. New York* (119 U. S. 110), *Nutting v. Massachusetts* (183 U. S. 533) and *National Union Fire Insurance Co. v. Wanberg* (260 U. S. 71).

³⁴ 322 U. S. 533 (1944).

³⁵ It is interesting to note that the state regulatory authorities were opposed to the majority opinion. Thirty-five state insurance departments joined in an *amicus curiae* brief opposing the classification of insurance as commerce. The dissenting opinions of Mr. Justice Jackson and Mr. Justice Stone indicated that the court's action might not only allow federal antitrust statutes to be applied to the insurance industry, but actually nullify the authority of individual states to regulate and tax the insurance industry.

³⁶ New York State Insurance Department [21], p. 69.

³⁷ Public Law 15, 79th Congress; 59 Stat. 33-34 (1945).

to be applied to the insurance industry only to the extent that the insurance business was not regulated by the states.

Exactly what was meant by regulation by the states was not made clear by the act nor by its legislative history.³⁸ The National Association of Insurance Commissioners, acting together with industry representatives, drafted two model bills which would establish state regulation of the business of insurance and preclude the application of the antitrust statutes against the insurance industry:

The overriding concern of the framers of these All-Industry model bills was to preserve the business and regulatory *status quo* and to demonstrate that rate-making, in particular, bureau rate-making, would be quite explicitly 'regulated' by the states. This approach was designed to provide a state regulatory umbrella under which cooperative rate-making by bureaus would be exempt from the Federal antitrust laws.³⁹

Eventually 44 states enacted new laws or amended existing laws to conform to the NAIC-All Industry model bills. These laws, known as prior approval laws, were the predominant form of regulation of the property-liability insurance industry from 1940 to 1970.⁴⁰

□ **Setting insurance rates under the prior approval regulatory system.**

Setting rates for property and liability insurance is a relatively complex procedure. This section describes the general technique and rate-making formulas which regulatory agencies have been applying in the "prior approval" states since 1944.

Under the prior approval laws fire insurance rate making involved the use of a fairly standard formula for the production and evaluation of bureau rates, rate revisions and rate deviations. Prior approval regulation does not involve the traditional public utility concept of fair rate of return on capital. Instead rates are established so as to yield a particular rate of return on sales (premiums). A standard rate of return on sales figure of 5 percent is employed in most states as a result of a recommendation by the National Association of Insurance Commissioners in 1921.⁴¹ This figure appears to have been picked out of thin air. It does not relate profitability to return on capital nor even to competitive profit margins on sales in other industries having the same capital base per unit of output or sales. In addition, investment income is not included as revenue. The technique appears to be *a priori* nonsensical. Although usually faced with many companies selling fire insurance in a particular state, adherence to bureau rates was so great that most regulatory agencies had to approve only bureau rate filings and a moderate number of rate

³⁸ Brainard and Dirlam [7], pp. 248-253.

³⁹ New York State Insurance Department [21], p. 72.

⁴⁰ Various other schemes were adopted in other states. California adopted a "No-Filing" system in 1947, essentially leaving rate determination to the market. Texas instituted state-made rates for many lines of insurance to which all insurers were required by law to adhere. North Carolina required bureau membership of all insurers. Other systems involved minor procedural differences from the prior approval laws.

⁴¹ It is interesting to notice that the NAIC 1921 profit formula bears a strange resemblance to a report written by Wilfred Kurth of the Home Insurance Company in 1919 and entitled "What Constitutes a Reasonable Underwriting Profit and the Method of Determining Same." The profit formula recommended by the insurance commissioners in 1921 appears in fact to be almost identical to that recommended by Kurth.

deviations based on the rating bureau filings themselves. In New York, for example, the rating formula was applied to the rate applications of the New York Fire Insurance Rating Organization in determining fire insurance rates.

In general, we can summarize the pricing formula for rates of a particular line of property insurance as one which sets insurance rates to cover losses, expenses and a profit factor (5 percent of premiums) based on aggregate experience for the industry as a whole in a particular state. While individual class by class rates are set in a similar way, aggregate profitability is the binding constraint. Rates set below the "formula" figure for some class will be compensated for by rates set above the "formula" figure for one or more other classes:

$$E + \frac{L_i}{P_i} + 0.05 = 1$$

$$P_i = \frac{L_i}{(1 - 0.05 - E)}$$

P_i = Total premiums for a particular territory and class,

E = Measure of historical operating and production expense ratio, and

L_i = Measure of historical losses (estimate of expected losses) for this territory and class.

If there are X_i identical risks in a particular territory and class (let us say they are automobiles), we would expect the average basic premium for this class and territory to be P_i/X_i per car. Letting $\Sigma P_i = P$ = total premiums indicated by the formula rating for all classes and territories in the state, we have:

$$P_i = \frac{L_i}{(0.95 - E)} \quad (1)$$

$$P = \frac{\Sigma L_i}{(0.95 - E)} \quad (2)$$

Note that if for some reason the insurance regulators feel that some particular class j yields an average premium $\frac{P_j}{K_j}$ which is in some sense "too high" (even though appropriate, given historical losses), the rating formula works in such a way that (2) is always met, whether or not (1) is met in all particular classes. This means that forced reductions in some classes' rates are automatically loaded on to the rates for the remaining classes of customers. We will return to this point when we discuss assigned risk pools, FAIR plans, and implicit attempts by regulatory agencies to redistribute income.

□ Price competition under the prior approval regulatory system.

Although the property-liability insurance industry possesses a market structure that is characteristic of a competitive market, the industry is notable for the limited amount of price competition which actually exists. Fire insurance companies only rarely tend to deviate from fixed bureau rates which have been approved by the regulatory authority. The automobile lines have tended to have a greater amount of price

competition, primarily because of the direct writers who often deviate below fixed bureau rates. But even by the end of the 1960s, a large proportion of automobile insurance was written at rates copied right out of the IRS guide. Most prior approval insurance laws authorized rating bureaus to make and file rates, rate changes, rating schedules, etc., for their member and subscriber companies. Companies not wishing to use the bureau rates had two options open to them: they could file deviated rates for one or more classes of insurance or they could submit an independent filing.

A company wishing to deviate from the bureau rates would simply have to file an application with the Superintendent of Insurance requesting to write one or more lines of insurance at some amount (say 15 percent) less than the rates then in effect for the members and subscribers of the rating bureau. The deviating company had to justify its rate deviation application by showing that its reduced rates were justified by lower costs than for the industry as a whole. Since its costs were lower, such a company could reduce its rates and still earn an "adequate" profit. Regulatory scrutiny of rate deviations appears to have been most concerned with keeping firms from filing deviations which would result in rates which were "too low" and might result in "destructive" competition.

Independent filings represent the submission of a complete set of rates, rating schedules, etc., which a particular company plans to follow, as opposed to the much simpler across-the-board reduction in bureau rates. For reasons which will become obvious presently, true independent filings have been rare in the insurance industry.

Observers have often criticized the administration of deviation procedures as tending to discourage companies from filing rate deviations. In many states rate deviations had to be renewed annually and rating organizations were often recognized as aggrieved parties. This allowed them to challenge rate deviations and independent filings. Groups of companies, acting through the rating bureaus, could thus challenge deviations and independent filings leading to protracted and often costly proceedings.⁴² The possibility of opposition to deviations and independent filings has been pointed to as being especially discouraging to new entrants and small firms wishing to sell at off-bureau rates.⁴³ Some of the more restrictive deviation procedures were gradually liberalized. The right of bureaus to act as aggrieved parties was banned in many states and the requirement that deviation filings be renewed each year was eliminated. In addition, many insurance departments continued to support companies which could justify lower rates whether they were independent filings or deviations from bureau rates.⁴⁴

Although it has been argued that price competition gradually increased over time under the prior approval laws,⁴⁵ it is extremely difficult to obtain adequate data necessary to quantify this trend in prior approval states. Some suggestive data for New York for the late 1960s are available, however.

⁴² U. S. Senate, Subcommittee on Antitrust and Monopoly [31], Part 2, pp. 924-940, and Part 3, pp. 1794-1845.

⁴³ *Ibid.*, pp. 938-939, 1129, 1179.

⁴⁴ *Cullen v. Bohlinger* 350 U. S. 803 (1955), and *Cullen v. Holtz and Allstate Insurance Co.* 6 N. Y. 2nd 1971 (1959).

⁴⁵ Brainard and Dirlam [7], pp. 243-244.

Table 2 indicates the overall price competition for the property-liability insurance in New York State.⁴⁶ Given the large number of companies in New York and the encouragement that the insurance department gave to price competition under the strictures of prior approval, the New York experience should indicate the maximum tendencies to price at other than bureau rates under the prior approval system. Fire insurance had only 8.9 percent of its premiums written at off-bureau rates while auto physical damage had 24.7 percent of the premiums written at off-bureau rates. Even in 1967, after many of the initial blockages to deviations and independent filings had been removed, the vast majority of the insurance policies were sold at the agreed upon cartel rate.

Table 3 indicates the results for the 30 largest insurers in New York State for three years and for three different lines of insurance. These figures differ from those of the first table because if a company filed a substantial number of deviations, all insurance it wrote was counted as being written at deviated rates, even if some of the rates were the same as the bureau rates. These calculations therefore count as a "competitive" price those rates or classes for which deviating and independent companies felt the bureau rate was justified by expense

TABLE 2
PERCENTAGE OF DIRECT
PREMIUMS WRITTEN IN NEW
YORK AT OFF-BUREAU
RATES - 1967

FIRE AND ALLIED LINES, AND EXTENDED COVERAGE	8.9%
HOMEOWNERS	22.5%
COMMERCIAL MULTIPERIL	14.4%
AUTOMOBILE LIABILITY	22.6%
AUTOMOBILE PHYSICAL DAMAGE	24.7%
LIABILITY OTHER THAN AUTO	23.1%
SOURCE: NEW YORK STATE INSURANCE DEPARTMENT [21], p. 93.	

TABLE 3
PRICE COMPETITION IN NEW YORK STATE AMONG THE TOP
30 INSURANCE COMPANIES AND ASSIGNED RISK PLAN
(UNDER PRIOR APPROVAL)

TYPE OF INSURANCE	% AT BUREAU RATES	% OFF BUREAU RATES	NO OF COMPANIES OFF BUREAU RATES
AUTO LIABILITY 1967	61.8	38.2	8
1968	61.3	38.7	8
1969	61.7	38.3	7
AUTO PHYSICAL DAMAGE 1967	53.1	46.9	8
1968	54.1	45.9	8
1969	53.0	47.0	7
HOMEOWNERS INSURANCE 1967	77.2	22.8	6
1968	77.9	22.1	6
1969	75.3	24.7	6
SOURCE: NEW YORK STATE INSURANCE DEPARTMENT [18], pp. 21-25 NEW YORK STATE INSURANCE DEPARTMENT [19], p. 63			

and loss experience. As a result, these figures indicate substantially more competition in these lines than the latter. Even so, less than a third of the companies were filing deviating or independent rates, and both the number of companies and proportions of premiums written at off-bureau rates remained approximately constant for the three years.

Of course there is the chance that the rating bureaus are setting prices at or near competitive levels, and this is why there have been so little price diversity and differences from bureau rates. Preliminary evidence indicates that this is not the case and that bureau rate making under prior approval rate regulation leads to substantially

⁴⁶ Under prior approval rate regulation.

TABLE 4
PERCENTAGE OF DIRECT
PREMIUMS WRITTEN AT OFF
BUREAU RATES IN CALIFORNIA

TYPE OF INSURANCE	
FIRE AND EXTENDED COVERAGE	32.0% (1967)
HOMEOWNERS	67.7% (1967)
AUTOMOBILE LIABILITY	49.9% (1966)
AUTOMOBILE PROPERTY DAMAGE	45.9% (1966)
SOURCE. NEW YORK STATE INSURANCE DEPARTMENT [21], p. 93	

less competition. The evidence takes two forms. One is the experience in California where regulation was not based on the all-industry prior approval system, but has been instead much closer to open competition in rate setting. The second source is New York State experience since January 1, 1970, when an experimental law went into effect, which essentially eliminated the prior approval system and substituted a California-type open competition ("no filing") rating system. Tables 4 and 5 should be compared with Tables 2 and 3.

Table 4 shows that at the same period of time, California, operating under an open competition law, had a substantially larger proportion of premiums written at off-bureau rates than had New York, which was operating under the prior approval regulatory statutes. However, Table 5 indicates that when the prior approval restrictions were eliminated, the power of rating bureaus reduced, and price competition encouraged in New York in 1970, the percentage of premiums written (implicitly the number of companies writing at off-bureau rates) at off-bureau rates increased substantially.

TABLE 5
COMPANIES WRITING PREMIUMS AT OFF BUREAU RATES
AMONG 30 TOP INSURERS AND THE ASSIGNED RISK PLAN
(NEW YORK 1970-1972)

TYPE OF INSURANCE	% OF PREMIUMS AT BUREAU RATES	% OF PREMIUMS AT OFF BUREAU RATES	NO. OF COMPANIES OFF BUREAU RATES
AUTO LIABILITY 1970	66.8	33.2	7
1971	66.7	33.3	6
1972	49.4	50.6	12
AUTO PHYSICAL DAMAGE 1970	56.0	44.0	9
1971	40.5	59.5	15
1972	14.6	85.4	26
HOMEOWNERS 1970	69.1	30.9	7
1971	42.8	57.2	10
1972	21.1	78.9	17
SOURCE NEW YORK STATE INSURANCE DEPARTMENT [18], pp. 21-25. NEW YORK STATE INSURANCE DEPARTMENT [19], p. 63			

The evidence supports the hypothesis that rate making in concert, combined with prior approval rate regulation, tended to discourage price competition. The rating bureaus with the knowing or unknowing help of the regulators had managed to maintain at least moderate cartel control of insurance prices in almost all of the property-liability insurance lines.

4. The performance
of the property and
liability insurance
industry

■ The notion of industry performance is of course quite ambiguous, primarily because of its multidimensional nature. This section will analyze the following three aspects of the performance of the property-liability insurance industry:

- (1) the efficiency of the prevailing production and distribution system;

- (2) supply shortages; and
- (3) profitability and capacity utilization.

The analysis is based on the notion that ideally we would like an insurance industry which provided insurance as cheaply as possible, confronting consumers with prices equal to the marginal (expected) costs of coverage and equilibrating supply and demand at the prevailing market price.

It is argued here that the combination of cartel rate making and other collusive behavior with state regulation of rates, rating classes, territories and other insurance practices has resulted in:

- (1) the use of a grossly inefficient sales and distribution technique (the American Agency System);
- (2) severe supply shortages; and
- (3) unnecessarily high prices, excess capacity, but probably only normal profits for the industry as a whole.

□ **Sales distribution technique: the American agency system.** Historically, fire and property insurance was sold through independent agents and brokers to the public. Insurance companies themselves did not possess their own retail sales personnel. For this service agents were paid on a fixed commission basis, with the commission set at some percentage of the customer's premium. Agents' organizations and rating bureaus often worked together to keep rates high by refusing to allow agents to sell the insurance of nonbureau firms. In return the agents sought high commission rates from the companies that they represented. Casual empiricism indicates that the selling costs through the American Agency System were higher than necessary to provide effective service. Mutual companies, writing through their own groups of agents or directly to consumers, often sold insurance at lower prices than the agency stock companies and still managed to have better underwriting results than the stock companies. The growth of direct writers in auto insurance after World War II, their ability to charge lower rates, and their financial success and fantastic growth rates reinforced the idea that the middlemen agents could be either eliminated or their numbers greatly reduced, resulting in substantial cost savings.

The analysis presented below indicates that the cost savings from full use of the direct writing or exclusive agency technique would result in substantial cost savings in the provision of insurance. It is argued that stock companies using the American Agency System have the highest costs and the highest prices and are employing an inefficient production technique, costing consumers hundreds of millions of dollars annually.

To explain company expense ratios we use the basic model employed in the returns to scale section. For the analysis presented in this section two relationships were estimated. In the first, observations on stock fire insurance companies, mutual auto insurance companies, and stock auto insurance companies were grouped together and the following relationship estimated:

$$E = \alpha + \beta \text{ DPREM} + \lambda \text{ INTER} + \gamma \text{ RATIO} \\ + \delta \text{ DWRITE} + \phi D_1 + \psi D_2,$$

where

$$\begin{aligned} D_1 &= 1 && \text{auto stock companies} \\ &= 0 && \text{otherwise} \\ D_2 &= 1 && \text{mutual auto companies} \\ &= 0 && \text{otherwise.} \end{aligned}$$

The estimate of δ is then used to measure potential efficiency from direct writing for the entire property-liability insurer. In the second, only observations on auto companies are used. The following relationship is estimated:

$$E = \alpha_a + \beta_a DPREM + \lambda_a INTER + \gamma_a RATIO + \delta_a DWRITE$$

The estimate of δ_a is used to estimate potential efficiency in the use of direct writing in auto insurance only. For the first relationship with observations on stock fire, mutual auto, and insurance companies, we obtain:

$$\begin{aligned} E = 33.85 - 0.65 DPREM + 0.73 INTER + 0.70 R & \quad (1. \\ (31.42) \quad (-0.73) \quad (0.80) & \\ - 10.82 DWRITE - 5.79 D_1 & \\ (-7.88) \quad (5.05) & \end{aligned}$$

$$R^2 = 0.51 \quad 157 \text{ observations.}$$

For the second relationship with observations only on auto companies we obtain:⁴⁷

$$\begin{aligned} E = 21.00 - 0.99 DPREM + 1.05 INTER + 7.90 R & \quad (4.31 \\ (10.84) \quad (-0.67) \quad (0.70) & \\ - 11.48 DWRITE & \\ (-9.65) & \end{aligned}$$

$$R^2 = 0.54 \quad 108 \text{ observations.}$$

The results indicate that overall, the expense ratios for direct writers average 10.82 percentage points less than the expense ratios for insurance companies *ceteris paribus*. For the auto companies themselves, the expense ratio is 11.48. In addition, the expense ratios for auto insurance are about 6 percentage points less than those for fire insurance. This latter result may have emerged because there has been more competition in auto insurance lines than in fire insurance. As a result, in order to stay competitive with direct writers, insurance companies have been forced to cut commissions to agents selling auto insurance. Finally (from ϕ_a) the cost structures for stock and mutual insurance companies are not significantly different from one another.

Let us assume that the technology for producing insurance contains only two activities: direct writing and indirect writing. Let us assume further that in a well-functioning insurance market the choice among these two activities would be based on which was the least cost activity (the outputs are the

⁴⁷ The results for the sample without the fire insurance companies (D_1 and D_2) were the following:

$$\begin{aligned} E = 21.20 - 0.99 DPREM + 1.05 INTER + 7.92 RATIO - 11. & \\ (11.53) \quad (-0.67) \quad (0.70) \quad (4.34) & \end{aligned}$$

$$R^2 = 0.54 \quad 108 \text{ observations.}$$

for the auto insurance industry only and then for the property-liability insurance as a whole. At current output and price levels we may calculate the efficiency gain from use of the direct writing system as follows:

$$EG = A \times B \times C,$$

where

EG = Efficiency gain per year

A = Expense saving (expressed as a percentage of premiums)

B = Total premiums

C = Proportion of premiums written by other than direct writers.

Factor (A) can be obtained from the regression results reported above and factor (B) is available from *Best's*. I was able to find a value for C for the $P - L$ industry as a whole (direct writers = 31.5 percent) but was not able to find a precise figure for the auto lines. I have estimated it as 50 percent direct writing:⁴⁸

$$EG (\text{Auto}) = \$830.0 \text{ million}$$

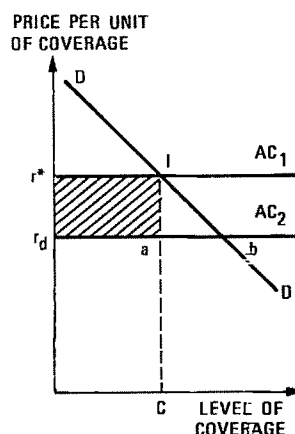
$$EG (\text{Total } P - L) = \$2.64 \text{ billion}$$

The measure can also be represented as the shaded area in Figure 1. Here r^* is the prevailing price per unit of coverage, and AC_1 and AC_2 represent the average costs per unit of coverage under the agency system and direct writing, respectively (including the "regulated" profit factor).⁴⁹

Now, this calculation assumes that the price remains at $r^* = AC_1$. However, even under regulation we would expect the price to fall to at least AC_2 , yielding an additional increase in consumer surplus equal to the small triangle labelled abl . If prices remain at r^* , the entire efficiency gain has been absorbed as higher profits to the insurance industry. The additional gain in efficiency from a price reduction will

FIGURE 1

SUPPLY AND DEMAND FOR INSURANCE COVERAGE



⁴⁸ This estimate is probably on the high side. I was able to obtain detailed evidence concerning the percentage of premiums written by direct writers in the State of New Jersey in 1971.

<u>Line</u>	<u>% of Premiums Written by Direct Writers</u>
Fire	8.24%
Homeowners	24.20%
Auto Liability	45.09%
Auto Physical Damage	49.69%
Total for 9 Property Insurance Lines	35.79%

⁴⁹ Let r^* = Price per dollar of coverage under the present system,

l_a = Expenses per dollar of coverage under present system,

L = Expected losses per dollar of coverage,

π = Profit per dollar of coverage,

l_d = Expenses per dollar of coverage under direct writing system,

C = Insured value,

$$r^* = l_a + L + \pi = AC_1,$$

$$r^*C = P,$$

P = Total premiums,

$$r_d = l_d + L + \pi = AC_2,$$

$$(r^* - r_d)C = EG = (l_a - l_d)C, \text{ and}$$

$$r = \frac{r^* + r_d}{2}$$

depend on the elasticity of demand and the difference between r^* and r_d .

Assuming a linear demand function⁵⁰ and Δr and ΔC small, we may approximate the dead weight loss resulting from setting price at AC_1 as

$$w = \frac{1}{2} (\text{Premium Volume}) (\text{Elasticity of Demand}) (\% \text{ Change in Expense Ratio})^2.^{51}$$

TABLE 6

DEAD WEIGHT LOSSES FOR SELECTED ASSUMPTIONS ABOUT DIRECT WRITING POSSIBILITIES AND THE ELASTICITY OF DEMAND FOR INSURANCE (\$ MILLIONS)

ELASTICITY	UNDERWRITING POSSIBILITIES			
	(1)	(2)	(3)	(4)
0.5	46.9	23.5	104.1	71.3
1.0	93.8	46.9	208.3	142.7
1.5	140.8	70.4	312.4	214.0
PRIMARY* EFFICIENCY GAIN	830.0	830.0	2,640.0	2,640.0
*AS CALCULATED PREVIOUSLY				

The dead-weight welfare gain from direct writing as shown in Table 6 is calculated for three different elasticity measures— $\eta = 0.5$, 1.0, and 1.5—based on different assumptions:

- (1) Only auto insurance can be successfully sold using direct writing and *all* companies charge r^* under the current system;
- (2) Only auto insurance can be successfully sold using direct writing and direct writers charge r_d under the current system;
- (3) All property-liability insurance can be successfully sold using direct writing and direct writers charge r^* under the current system;
- (4) All property-liability insurance can be successfully sold using direct writing and direct writers charge r_d under the current system.

⁵⁰ This ignores income effects.

⁵¹ Under these assumptions the dead-weight welfare loss may be approximated as

$$w = \frac{1}{2} (r^* C) \eta \frac{\Delta r^2}{r},$$

where

$r^* C$ = total premiums (P),

η = elasticity of demand, and

Δr = change in the price per unit of coverage ($r^* - r_d$).

But what is the relationship between the percentage change in the price per unit of coverage and the expense ratio reduction which we have already calculated? We know that

$$r^* C = P$$

and

$$r^* - r_d = l - l_d$$

so that

$$C(r^* - r_d) = C(l - l_d) = \text{total expenses}$$

and

$$\frac{C(l - l_d)}{P} = \Delta E \quad (\Delta E = \text{difference in expense ratios}).$$

These calculations give us some feeling for the range of values for the welfare gains which society might achieve by moving from a system of independent agents to one of direct writing. The minimum efficiency gain is \$830 million dollars per year under the assumption that the auto insurance market is the only one congenial to direct writing and that the elasticity of demand for coverage is zero. The maximum gain is \$2.64 billion under the assumption that the entire property-liability industry is susceptible to direct writing and that the elasticity of demand for coverage is 1.5. For argument's sake we shall say that the efficiency loss is \$1.5 billion per year because of the prevalence of the American Agency System.

The natural question which arises here is why the American Agency System continues to exist at all. If direct writing is cheaper, and if direct writers can justify lower rates through deviation procedures before state regulatory commissions, why does the direct writing sales technique not drive the American Agency System out of the game? We turn to this question now.

A possible answer to this question is that the system of small independent agents dealing with the public is in some sense "better." The argument goes that the agent can service the specific problems and needs of each of his individual customers. For some lines of insurance and for certain types of properties with special insurance problems there certainly is need for individual insurance counseling. However, the Senate Judiciary Subcommittee hearings are full of testimony that this is just the kind of thing which the small independent agent does poorly. He has neither the training nor the day-to-day experience to handle tricky insurance problems. The customer in need of special advice will either have to deal directly with a service representative of a company or through a large agency or independent consultant with personnel capable of handling special problems. The kinds of things which the small independent agents handle well are the standard recurring day-to-day insurance coverage applications. Most auto insurance is of this form; standardized policies, printed application forms, and a book with rates printed in it. The customer need only appear, specify his age, sex, driving record, type of car, and desired coverage, and his policy application is complete. The agent will then "try" to get the customer coverage. This is also just the type of thing which direct writers do very well. For many types of homeowners policies and straight fire policies (especially for small dwellings and commercial properties) as well as auto insurance, the agent often does little more than fill out a pre-printed form and act as a go-between for the customer and the insurance company. In many urban areas the assigned risk plans (auto) and the FAIR plans (fire, extended coverage) are really no more than direct writing organizations for the insurance industry as a whole, serving those persons who cannot get insurance in the voluntary market. While the agent continues to be a middleman in this situation, he is certainly redundant. Much of this business could easily be handled by direct mail to customers, with service personnel available for questions or through large area sales outlets.

The advantages of independent agent contact for many types of insurance customers appear to be small or even perverse. Since an insurance company which integrates forward into sales can both sell insurance by mail and maintain regional offices to handle claims

and deal with special problems, the need for the independent agent *per se* appears to be nonexistent.⁵² Why then, in the face of large cost savings, have many companies not integrated forward into sales to eliminate the inefficient agency system as a sales technique?

As a matter of fact, no major company has ever switched over from the agency system to the direct writing system. Direct writers in existence appear to have begun as direct writers. While it would seem reasonable for companies to switch from an agency operation to a direct writing operation, there are substantial blockages to this. One of the primary barriers is legal. Since the National Fire Insurance case (1904), the "property rights" to an insurance customer belong to the agent who brought the customer to the company. A company is legally prohibited from going out and soliciting the business of one of its customers itself. It must deal with the customer through the agent who generated the business. Therefore, a company which wished to switch from a system of generating business through independent agents to direct writing would essentially have to give up all his existing customers. A large company might attempt to make this transition region by region or state by state, funding the change-over costs through remaining agency business; however, the strong national agents' trade associations would certainly frown on this. A company trying to shift gradually to direct writing might find itself informally blacklisted by the trade associations and lose its agents everywhere. In addition, evidence cited above indicates that there are moderate to high entry barriers facing potential entrants into the direct writing market. As a result, although the property-liability insurance industry taken as a whole possesses concentration ratios indicating a competitive market structure, the direct writing segment appears to be much more oligopolistic.

The problem of consumer information has also probably been a contributing factor. Undoubtedly, there is no other product for which consumer ignorance is so prevalent. Many consumers are unaware that there are price differences among insurance companies and that it is not necessary to go through an independent agent to obtain insurance. Comparative price shopping is very difficult since price differences for comparable coverages are not readily available in printed form and because it is difficult to obtain information from friends and neighbors. I can get good information on price differences and price levels for all kinds of products from the fellow in the next office if he has purchased these products and/or shopped around for them. Asking him about his insurance is of little value since he is in a different risk class, lives in a different community, and drives a different kind of car. In many states deviations from bureau rates were so rare that many consumers simply believed that all rates were the same and just left it to their agent to find them a company. The agent himself had little incentive to obtain cut-rate insurance since his commission usually varied directly with the size of the premium.⁵³ In the presence

⁵² By independent agent we mean the agent whose incomes from commissions on policies he can actually sell. There is no reason to believe that truly independent insurance consultants, charging fees directly to customers with insurance problems, would not thrive with the elimination of the agency system. If services are in fact needed beyond those provided by the insurance firm, we would expect such consultants to be developed. The choice would then be left to consumers, and fees charged would not necessarily be tied to premium volume generated.

⁵³ This is less of a problem for large customers since there is some competition

of such large information problems it is not surprising that it has taken so long for the direct writers charging lower rates to capture a sizable portion of the property insurance markets, even given their relatively strict underwriting policies.

Regulatory commissions have given only limited recognition to this consumer information problem. Under the prior approval system, regulation seems more concerned with making sure that rates were "adequate" than with encouraging consumers to take advantage of lower rates where available. Only recently have the Pennsylvania and New York insurance departments published guides which give consumers more information on price differences among companies, for different classes of insurance, and in different areas of the state. It was recognized in New York that such information was essential if the new open competition rating system was to work in a way which would promote "adequate" and not excessive insurance prices through competitive market forces.

Even with these barriers, however, it would not be unreasonable to suppose that the large cost advantage direct writers possess would allow them to reduce prices so much that they could increase their market shares very quickly. Although direct writers have been increasing their share of the market, the movement has been gradual. Almost nonexistent before the Second World War, the four largest direct writers possessed 15.5 percent of the total auto insurance market in 1954 and 25.5 percent of the auto insurance market by 1971.⁵⁴

By and large it does not seem as if the direct writers have tried to take over the market rapidly. Although the direct writers and many mutual companies tend to file across the board 10-20 percent deviations on all lines of insurance, they also tend to have very strict underwriting policies. They attempt to take the "cream" of each of the risk classes. Since the risk classes are not really homogeneous in terms of type of person or neighborhood character, the direct writers tend to look to additional characteristics which will differentiate out the best risks in the class. Driving record, accident record, home address, occupation, etc., are the kinds of things direct writers and some mutual companies look to in obtaining the "quality" risks in each class. The sales situation appears to be one in which the low expense direct writers charge reduced rates for quality risks, leaving the rest of the market to be serviced by agency companies and assigned risk plans.

While this behavior may appear to be paradoxical, we shall see below that it is not. On the contrary it appears to be consistent with profit-maximizing behavior of low cost oligopoly firms (the direct writers) insulated from entry and faced with short-run "capacity" constraints.

□ **Supply shortages.** The problem of supply shortages in property and liability insurance has resulted in a great deal of criticism of the industry. By supply shortages we mean that customers who desire

among agents for customers' business. Even here severe information imperfections probably keep the market from working very effectively.

⁵⁴ It should not be assumed that direct writing is only suitable for auto insurance. The top 4 direct writers wrote 16.4 percent of homeowners' premiums in New York State in 1971 up from 12.4 percent in 1967.

insurance at prevailing market rates simply cannot obtain it through standard channels or cannot get so much insurance as they would like at posted rates. Fire, extended coverage, and theft insurance became very difficult to obtain in the ghetto core areas of many cities in the United States in the 1960s.⁵⁵ The exact extent of the shortage is difficult to measure, but some estimates run as high as 80 percent nonavailability in particular ghetto areas. Fire insurance pools, known as FAIR plans, have been set up in many states to provide insurance to those who could not get it in the open market (often at open market rates). Similar types of shortage problems have arisen in automobile insurance as well. These problems became especially evident when many states initiated mandatory financial responsibility laws, resulting in a situation in which those who could not get insurance in the voluntary market could no longer obtain car registrations.

Assigned risk plans were established in all states to provide at least the basic minimum insurance coverage to everyone (again often at bureau rates). The exact operations of the FAIR plans and the assigned risk plans will not be discussed here. However, FAIR plans in many states have been forced to write substantial volumes of fire, extended coverage, and burglary insurance. The New York FAIR plan was, by 1972, the largest fire insurance underwriter in New York state. The supply shortage situation in automobiles was not so severe, however. An examination of the activities of New York State's Assigned Risk Plan is revealing.⁵⁶

Except for the years following the introduction of compulsory automobile liability insurance, the proportion of vehicles unable to obtain insurance in the voluntary market hovered near 10 percent and was 12.7 percent in 1971.⁵⁷ The situation in ghetto areas was

⁵⁵ *Meeting the Insurance Crises of Our Cities* [26].

⁵⁶

PERCENTAGE OF TOTAL CAR-YEARS INSURED THROUGH
THE ASSIGNED RISK PLAN IN NEW YORK

<i>Year</i>	<i>Percentage</i>
1957	3.1
1958	6.7
1959	9.2
1960	10.4
1961	10.0
1962	9.3
1963	8.1
1964	7.8
1965	7.4
1966	7.7
1967	8.2
1968	8.4
1969	8.8
1970	9.9
1971	12.7

Source: New York State Insurance Department [17], p. 16.

⁵⁷ There may be an element of perverse behavior here which is putting some people into the assigned risk plan who might be able to find insurance in the voluntary market. This seems to be related to the American Agency System. Before 1969, agents in New York received only a 5-percent commission on business written through the assigned risk plan, based only on the basic premium charge which at that time was identical to the bureau rates in the voluntary

considerably worse than the yearly averages. In Bronx County South fully 34 percent of the automobiles were insured through the assigned risk plan in 1971.⁵⁸ Youthful owners or principal operators also found it difficult to obtain automobile insurance in the voluntary market, with 29 percent of the vehicles registered to this class of operators insured through the assigned risk plan in New York State. The proportion ran as high as 48 percent in one particular rating territory.⁵⁹

In addition, until recently most states only provided minimum coverage through the assigned risk plans so that many of those in the plan may not have been obtaining the quantity of coverage which they desired at prevailing prices. In fire and extended coverage as well as in the auto insurance lines a far from insignificant proportion of the population is not being serviced adequately by the voluntary market.

Clearly an understanding of why supply shortages exist is crucial to the evaluation of the performance of the insurance industry under regulation. The contention here is that supply shortages exist because regulatory authorities have refused to allow the creation of enough truly homogeneous risk classes. Rates determined on the basis of average historical losses for a particular class may be profitable for a "typical" risk. However, as long as risk classes are not homogeneous from the viewpoint of insurers, and insurance underwriters can differentiate between "good" risks and "bad" risks within a particular class, no risk will be treated as "typical." On the contrary most insurance companies will attempt to insure the "good" risks first and may or may not insure the bad risks.

Proposition 1: The existence of supply shortages within a particular heterogeneous class will depend on the expected profits associated with its identifiable subclasses and the ability of insurers to differentiate successfully between good and bad risks within the class.

Proposition 2: Supply shortages within a particular heterogeneous rating class will vary inversely with the underwriting profit for the "typical" risk in the rating class.

market. Since commission rates on business written through the voluntary market were substantially higher, agents had an incentive to write their business through the voluntary market if they could. In 1969 the insurance law was amended to allow agents to charge 5 percent not only on the basic premium, but on the basic premium plus surcharges made due to accident and driving records. These surcharges could increase premiums as much as 100 percent, and although surcharge provisions were available to companies in the voluntary market, they were rarely used (especially for minor traffic violations). As a result, agents could now make more money by getting insurance for their clients at a surcharged rate through the assigned risk plan than they could by getting insurance at standard rates in the voluntary market. This incentive was strengthened when rates in the assigned risk plan were allowed to go above the standard bureau rates in 1971 and when the actual rates charged by many companies under the new open competition fell below bureau rates. Many agents were faced with a situation in which 5 percent of a large premium looked much better than 10 percent of a much smaller premium. They had an incentive to feed business into the assigned risk plan which had no real way of verifying whether a risk was really uninsurable in the voluntary market. This is probably why the number of risks in the assigned risk plan increased so dramatically between 1970 and 1971.

⁵⁸ New York State Insurance Department [17], p. 19.

⁵⁹ *Ibid.*, p. 22.

Arguments in favor of these two propositions are presented in the context of a simple model of the *short-run* underwriting behavior of a profit-maximizing property insurance company.

Consider a typical insurance company with some fixed capital base (K) “backing” its insurance operations. The company earns income *from its insurance operations* in two ways:

- (1) Direct underwriting profits; and
- (2) Investment income from prepaid premiums, often called “unearned premium reserves.”

Let us assume for simplicity that the typical firm sells one type of insurance to only one class of customers.

Let π^* = profit per premium dollar for the “average” or typical customer in the class. π^* is determined by the regulatory authority and defines a price per unit of coverage for *all* customers in the class:

$$\pi^* = 1 - l - e,$$

where

l = Expected losses per premium dollar for the class as a whole

e = Expenses per premium dollar.

Let us assume further that there are two types of risks in this class—“good” risks and “bad” risks—defined by some indicator variable not used for determining the rating class itself. Let us say that this variable is dichotomous and indicates “good” neighborhoods and “bad” neighborhoods within the territory used for establishing the prevailing price for all risks.

Let

π_B = Expected profit per premium dollar for the “bad” risks in the group;

π_G = Expected profit per premium dollar for the “good” risks in the class;

$\pi_B < \pi^* < \pi_G$, and

$y_i = \pi_i + r^{**}$,

where

y_i = Expected net income from writing a risk with characteristic i , and

r^{**} = Investment return per premium dollar from the “unearned” premium reserve.

Short-run expected profit-maximizing behavior of insurance firms, given π_G , π_B , and r^{**} , implies that an insurance company will write all subclasses of the heterogeneous rating class for which

$$y_i \geq 0$$

and will not write subclasses for which

$$y_i < 0.$$

In general the higher is π^* (*ceteris paribus*), the more subclasses (with progressively poorer loss expectations) will be written. In the two-subclass case, the higher is π^* , the more likely the bad risk group will be taken on. If $\pi_B + r^{**} \geq 0$, both subclasses will be written. We

therefore expect that a class will still be written for some values of $\pi_i < 0$.

This simple model includes only two types of risks and implies a simple dichotomous decision index which specifies good neighborhoods and bad neighborhoods, which in turn determines "good" risks and "bad" risks (and the expected profitability of each). Especially for automobile insurance, we might expect to find a decision index which takes on many values and is a function of such things as neighborhood residence, driving record, accident record, age, etc., which have not been used in defining the rating class itself. We define the index so that it varies inversely with the quality of the risk.

Let

$\pi_I = F(I)$ (where I is the value of the decision index and π_I are expected profits),

$F'(I) < 0$.

In the short run we would expect firms to write all customers for which $F(I) \geq -r^{**}$ which is essentially the result for the more simple case. Quite simply firms just will not write risks for which they expect to lose money. If risk classes are in fact really heterogeneous, all risks will be insured only if π_i for the "worst" identifiable group in the class is equal to or greater than $-r^{**}$. This means, however, that persons in the intramarginal classes will be paying too much for insurance. In the short run firms would be earning profits in excess of what would induce them to write each subclass if each were priced according to its own expected losses in a competitive market with truly homogeneous risk classes. Note finally that π^* may be so low that nobody within a particular class will be insured.

There is substantial evidence to indicate that rating classifications have been too large in many states. In New York City, for example, although there are 115 building hazard classifications, rates for each are based on data from a very large territory—the entire City of New York. Companies often believed that within any hazard class, some neighborhoods have higher expected losses than others. A system of "redlining," where some areas of the city were designated as not insurable, developed. The number of areas designated seems to have increased during the late 1960s as underwriting profits (π^*) declined and expectations of losses increased. The large supply shortages which resulted in many major cities appear to have been caused because companies felt that the premium rates did not justify the expected returns from writing in these areas. Attempts by insurance companies to put surcharges on particular neighborhoods have generally been resisted by regulatory authorities as being discriminatory. This was at least partially the fault of the industry, which did not collect industrywide loss data by neighborhood. Even if it had, changing expectations based on things other than historical loss experience might still have resulted in "regulated" prices which were too low for some risk classes.

Similar behavior appears to exist for auto insurance. While the number of territory classifications for auto insurance in New York City is much larger than for fire insurance, there is evidence that insurers tend to avoid people with poor driving records or poor traffic

records, people in certain occupations, aliens, etc.⁶⁰ In particular, new inexperienced drivers tend to be avoided by insurance companies. For some of these characteristics surcharges are available, for others they simply are not.

The behavior of insurance companies in "risk selection" certainly appears to have been justified, when we look at the loss experience for auto risks insured through the voluntary market and those insured through the assigned risk plan. The losses and loss adjustment figures for assigned risk drivers are much higher than those for drivers insured through the voluntary market.⁶¹ In 1970 the ratio of losses to premiums for risks insured through the voluntary market was 0.68, while it was almost twice as high as 1.28 for assigned risk drivers. Similar results were obtained during the previous 10 years.

Given excess demand and the contention by underwriters that the excess demand exists because particular subgroups of existing rate classifications are identifiably "unprofitable" at bureau rates, why does the regulatory authority not allow more rate classifications? In many cases *they have* allowed more rate classifications and special surcharges. Convincing regulatory authorities that additional risk classifications are justified is not easy, however. Data must be collected justifying new classifications, and often must be of at least five years' duration to be used for making rates. Chicago was recently broken up into four rating areas for auto insurance and New York City will soon be rated by borough for fire insurance. The burden of proof for justifying new classifications is up to the companies in the industry and must be supported by extensive loss data.

In addition, regulatory authorities have resisted attempts at certain types of classification schemes on equity and distributional grounds. The fact that the poor, the nonwhite, and the young would tend to pay more for all types of insurance, if risk classes were truly homogeneous, has certainly had a lot to do with the policies of the regulatory authorities. Insurance regulators seem to have felt that charging higher rates in slum areas, to black people, to persons with particular occupations, etc., for what to the layman appears to be an identical product, could present severe political difficulties.⁶² The true price of

⁶⁰ See New York State Insurance Department [17], p. 36.

⁶¹

NEW YORK STATE AUTOMOBILE LIABILITY INSURANCE
PRIVATE PASSENGER CARS
(All Companies 1960 to 1970)

<u>Accident Year</u>	<u>Loss and Loss Adjustment Ratio</u>	
	<u>voluntary market</u>	<u>assigned risk plan</u>
1960	0.58	1.15
1961	0.55	1.07
1962	0.57	1.12
1963	0.60	1.14
1964	0.64	1.23
1965	0.66	1.30
1966	0.64	1.27
1967	0.67	1.28
1968	0.73	1.39
1969	0.72	1.39
1970	0.68	1.28

Source: New York State Insurance Department [17], p. 29.

⁶² Attempts to separate out high risk areas have often been met with stiff

insuring some bad risks profitably would be so high that they appear to be "unconscionable" to many concerned. By lumping good risks together with bad risks in identical rating classifications, regulators could approve posted rates which appeared to be fair for all rating classifications. However, true cross-subsidization would be impossible because insurance companies are not forced to provide service to all customers who demand it.⁶³

A cross-subsidy scheme which probably would have worked in the traditional public utility setting could not work here so long as companies could easily identify the poor risks in each class which would not be profitable to underwrite. The result would not be cross-subsidization, but a lack of supply availability at posted prices for many groups of consumers.

The introduction of assigned risk plans and FAIR plans was an ideal solution for the regulatory authorities. Companies were now being forced, indirectly, to provide all customers with at least some service. Since this service was often provided at bureau rates established on the basis of loss experience for customers insured both inside and outside of assigned risk plans, good risks ended up subsidizing bad risks. It has been estimated that this subsidy amounts to an increase of \$9 per car per year on the rates of auto liability risks insured through the voluntary market, and a reduction of \$76 per car per year on the rates of assigned risk participants below what would be required to cover losses, expenses, and a 5-percent profit factor.⁶⁴ The insurance industry appears to have been satisfied with this arrangement as long as overall rates were high enough to keep the number of customers in special insurance pools small and overall profitability acceptable. Recent experience with poor industry profits appears to be making the insurance industry as a whole far less satisfied with the current arrangement than it has been in the past. The question of profitability is one to which we turn now.

□ **Profitability.** Profitability, in the context of a "fair rate of return" on capital, is an important aspect of the pricing process in most regulated industries. Overall allowed rates of return are normally determined by regulatory agencies in rate hearings on a company by company basis and price structures yielding the allowed rate of return approved.⁶⁵ The question of excess profits and of profit measurement methodology was not very much of an issue in insurance rate regulation until recently. A standard formula, yielding "fair" profits to the industry as a whole, was agreed to by the National Association of Insurance Commissioners in 1921. It has been in effect in most regulatory jurisdictions ever since. Recent criticism of rising insurance rates, coupled with a reconsideration of the effectiveness of regulation, has led to an examination of both the industry's profit performance

resistance by the persons who live in those areas. Attempts to get surcharges put on the fire insurance rates in the Lower East Side area of New York City by the insurance raters were met with great resistance. The insurance department indicated that such surcharges would be "discriminatory" unless backed up by substantial loss data for the area in question.

⁶³ See Posner [25].

⁶⁴ U. S. Senate, Subcommittee on Antitrust and Monopoly [31], Part 18A, p. 12677.

⁶⁵ A notable exception to company by company rate determination was the FPC's area rate method of determining the wellhead price of natural gas.

and the profitability measures to be used in approving and “surveying” property-liability premium rates.

□ **Traditional measures of profitability.** The traditional measure of profitability used for property-liability insurance rate making is based on the “1921 Profit Formula” adopted by the National Association of Insurance Commissioners in 1921. While the rationale for this formula was based on the situation in fire insurance rate making, the basic profit allowance and its calculation have been applied to all major property-liability lines. This formula contains the following provisions:⁶⁶

- (1) Underwriting profit (or loss) is arrived at by deducting all incurred losses and incurred expenses from earned premiums;
- (2) No items of profit or loss connected with the so-called banking end of the business should be taken into consideration; and
- (3) A reasonable underwriting profit is 5 percent plus a 1-percent conflagration hazard (as modified by the NAIC in 1948) for fire insurance rate making.

In most regulatory jurisdictions the profit formula has been applied to the performance of the industry as a whole and not on a company by company basis. Individual companies wishing to reduce rates below approved bureau rates normally had to show that their operating expenses (but not their losses) were sufficiently below those for the industry as a whole, so that the rate reduction would still yield them the statutory profit figure. To evaluate this profitability measure and the actual profit performance of the insurance industry, the following simple model of the behavior of the insurance firm in the long run will be useful.

□ **A simple model of the behavior of a property insurance firm in the long run.** The insurance firm is conveniently conceptualized as a levered investment trust. It obtains capital from an initial sale of stock and invests most of this capital in securities (both stocks and bonds). This is the strict “investment trust” aspect of the firm. The firm’s portfolio is then used as backing for its insurance operations. The insurance operations yield two sources of net income: direct income from underwriting (premiums less losses and expenses) plus investable capital (unearned premium reserves) made up of prepaid premiums, with which the firm purchases more securities for its portfolio. For U. S. stock companies the value of capital attributable to unearned premium reserves is equal to about 80 percent of the equity capital (including retained earnings or surplus). By engaging in insurance operations, the “investment trust” gains additional investable capital and additional income, but must also assume the additional risk of engaging in the insurance business.

Let

r^* = Expected rate of return on equity portfolio;

I = Expected underwriting profit after losses and expenses per premium dollar (as determined by regulatory authority and rating bureau);

⁶⁶ *Proceedings of the National Association of Insurance Commissioners, 1922*, pp. 19–29.

P_i = Total premiums written by firm i ;
 r^{**} = Expected rate of return on "insurance capital" portfolio;⁶⁷
 K_i = Equity investment in firm i
 U_i = Unearned premium reserve possessed by firm i .

Total expected net income of the insurance firm in a given period is

$$\pi_i = r^* K_i + I P_i + r^{**} U_i, \quad (3)$$

and the rate of return on the operations of the firm as a whole is given by

$$r_I = \pi_i / K_i. \quad (4)$$

Assuming U_i/P_i remains constant at a value of approximately 0.8 we may rewrite (2) as

$$r_I = r^* + \frac{P_i}{K_i} [I + 0.8r^{**}].$$

With r^* , I , and r^{**} constants given by the market, the expected rate of return for the insurance firm varies directly with the ratio of premium volume to equity capital, *ceteris paribus*.

Let us assume that premium rates have been fixed at some level by the rating bureau and/or regulatory agency so as to yield I and that all firms adhere to these "cartel" rates. This premium rate level defines a level of total insurance coverage demanded, and in turn a level of total premium volume P ($P = \sum P_i$). If all firms are identical and there is free entry we would expect the long-run equilibrium positions of all firms to be characterized by:

$$r_I^c = r_I = r^* + I \frac{P_i}{K_i} + 0.8r^{**} \frac{P_i}{K_i}, \quad (5)$$

where r_I^c is the competitive opportunity cost of capital.

Then r_I^c is defined as follows:

$$\begin{aligned}
 r_I^c &= r_c + g\left(\frac{P_i}{K_i}\right), \\
 g'\left(\frac{P_i}{K_i}\right) &> 0, \\
 g''\left(\frac{P_i}{K_i}\right) &\geq 0, \quad r_c > r^*,^{68}
 \end{aligned}$$

where r_c is the expected return from operating a pure investment trust, and where the opportunity cost of capital rises as the investment portfolio "backs" more and more insurance claims (since prices are fixed here, this is represented by premium volume):

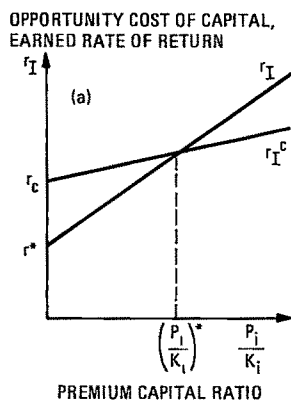
$$r_c + g\left(\frac{P_i}{K_i}\right) = r^* + I \left(\frac{P_i}{K_i}\right) + 0.8r^{**} \left(\frac{P_i}{K_i}\right), \quad (r_c > r^*). \quad (5')$$

⁶⁷ We differentiate between r^* and r^{**} , since certain regulatory restrictions may limit the ways in which certain sources of investable funds may be invested.

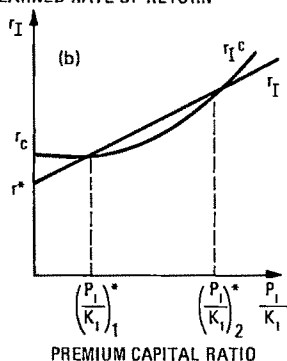
⁶⁸ Here $r_c > r^*$ because transformation of an investment trust into an insurance company, even if no insurance is actually written, requires some expenditures of funds for licensing and auditing (which would not otherwise be incurred) and also restricts somewhat the kinds of securities that may be purchased.

FIGURE 2

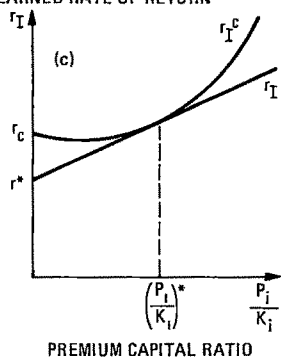
OPPORTUNITY COST OF CAPITAL
VERSUS EARNED RATE OF RETURN



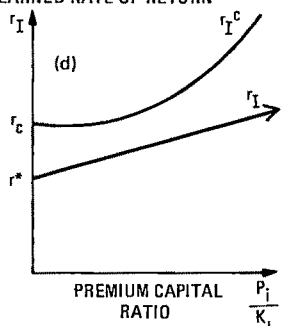
OPPORTUNITY COST OF CAPITAL,
EARNED RATE OF RETURN



OPPORTUNITY COST OF CAPITAL,
EARNED RATE OF RETURN



OPPORTUNITY COST OF CAPITAL,
EARNED RATE OF RETURN



Equation (5') may be solved for an equilibrium value of $\frac{P_i}{K_i}$.

Single and multiple equilibria are possible, depending upon the shape of $g(\cdot)$ and the level of I . The four possible situations are depicted in Figure 2 (a)–(d).

The force driving firms to an equilibrium point is the entry and exit of equity capital into and out of the industry. Since ΣP_i is fixed (since prices are fixed), adjustment takes place in terms of K :

- if $r_I > r_I^c$, capital will enter the industry;
- if $r_I < r_I^c$, capital will leave the industry; and
- if $r_I = r_I^c$, equilibrium is reached.

We may think of this entry in terms of additional identical firms (P_i for each firm then declines with entry, K_i remaining constant), as existing firms expanding their capital base (P_i constant and K_i increasing), or as a combination of both.

Case (a) where r_I^c is a linear function of $\left(\frac{P_i}{K_i}\right)$ leads to a unique and stable equilibrium position. Case (b) where r_I^c increases more than proportionately with $\left(\frac{P_i}{K_i}\right)$ yields two equilibrium positions, however only $\left(\frac{P_i}{K_i}\right)^*$ is stable. Firms can avoid moving beyond $\left(\frac{P_i}{K_i}\right)^*$ by refusing to supply all insurance demanded, so that the system is not explosive. Case (c) is a polar case of (b) and is stable as long as firms can refuse to write some insurance demanded or recognize that entry will drive down the opportunity cost of capital. Finally (d) is a situation in which income from insurance operations is just not worth the risk.

Let us concentrate on the equilibrium positions in (a) and (b) $\left[\left(\frac{P_i}{K_i}\right)^*\right]$. The equilibrium premium-capital ratio for individual

firms will vary inversely with I (*ceteris paribus*). In (a) any $\left(\frac{P_i}{K_i}\right)^*$ value can be supported at some value for I (and this is probably an unrealistic characterization of the cost of capital for this reason).⁶⁹

In (b) only a limited range of stable values for $\left(\frac{P_i}{K_i}\right)^*$ can be supported by varying I . In both cases there is some lower bound on I below which firms will simply withdraw from the industry. Call this floor value on the underwriting profit level I^* .

In this world then, by defining a particular value for $I \geq I^*$, the regulatory authority and the rating bureau are implicitly defining an equilibrium level of $\left(\frac{P_i}{K_i}\right)$ at which firms will sell insurance. The

higher I is, the smaller the value of $\frac{P_i}{K_i}$ will tend to be. Since there is easy entry, we would expect that the earned rate of return for individual firms and for the industry as a whole should be equal to the

opportunity cost of capital for all $I \geq I^*$. As long as the regulatory authority sets $I \geq I^*$, insurance will be provided in the long run. However, if the regulatory authority should also undertake to prescribe a maximum $\left(\frac{P_i}{K_i}\right)$ ratio at which firms must write insurance, it will either be applying an ineffective constraint—if it is greater than $\left(\frac{P_i}{K_i}\right)^*$ —or will force firms out of the market—if it is less than $\left(\frac{P_i}{K_i}\right)^*$. Proposals to have regulators set both I and $\left(\frac{P_i}{K_i}\right)^*$ may create considerable problems.

In a truly competitive market we would expect insurance premium rates to be driven down to the point where they were just high enough to clear the market. The competitive underwriting profit would therefore be I^* (which *could* be negative because of the earnings from the increased premium reserves) which would be just high enough to coax firms into the industry, providing insurance to all who demand it with each firm operating at some premium-equity capital ratio which we may call $\left(\frac{P_i}{K_i}\right)^{**}$. If cartel prices are set above I^* , market equilibrium will result in a $\left(\frac{P_i}{K_i}\right)^*$ value smaller than $\left(\frac{P_i}{K_i}\right)^{**}$. The cartel pricing situation will result in less real insurance coverage's being offered (assuming demand is not perfectly inelastic), more capital (capacity) per premium dollar and per unit of real insurance coverage, and a lower rate of return (although still equal to the opportunity cost of capital) than would be achieved in a competitive market. To the extent that consumers are indifferent among the $\left(\frac{P_i}{K_i}\right)$ values $\leq \left(\frac{P_i}{K_i}\right)^*$, they will be better off with a competitive market.

A possible rationale for the cartel arrangement thus presents itself. If the demand for insurance coverage has a price elasticity less than or equal to unity and if demand itself is growing only slowly, the introduction of competitive pricing necessarily means that some existing firms will be forced to exit from the industry. Even for some values of the price elasticity greater than unity, exit would be necessary. Therefore, from the viewpoint of existing firms in the industry, the elimination of cartel pricing may indeed have resulted in excess capacity and competition destructive to some existing firms. The fact that a cartel with free entry could not achieve excess profits in the long run for the constituent firms is not a deterrent to the establishment or maintenance of a cartel. With the cartel all existing firms would stay in the ball game. The elimination of the cartel would probably have meant that some of the existing firms would have been forced out of the market. Once a cartel-type equilibrium has been attained (or if there is excess capacity for some other reason), existing firms have an interest in maintaining (or creating) the cartel to keep themselves in the market.

We may conclude that one of the reasons many insurance companies (especially the larger ones) have not resisted the movement

to open competition in insurance pricing is that the cartel arrangement was falling apart with the direct writers and the eased deviation procedures. Effective prices were being driven toward the competitive level, and no doubt many of the larger companies felt that they could do much better, at least in the short run, with open competition, at the expense of many of the smaller companies.

Before moving on, certain consumer protection arguments in favor of regulated prices should be made. It is often argued that because of consumer ignorance, consumers appear to be indifferent between companies selling insurance with different $\left(\frac{P_i}{K_i}\right)$ values, but really would not be if they were apprised of the significance of the $\left(\frac{P_i}{K_i}\right)$ ratios to them. In some sense, the higher the value of the $\left(\frac{P_i}{K_i}\right)$ ratio of the firm from which one purchases an insurance policy *ceteris paribus*, the lower is the quality of the "product."⁷⁰ The quality of the product is reduced because the higher is $\left(\frac{P_i}{K_i}\right)$, the higher is the probability that the insurance firm will go bankrupt and that losses incurred will not be paid off. This argument implies that with perfect information, consumer demand would depend not only on price but also on the $\left(\frac{P_i}{K_i}\right)$ ratio maintained by firms selling contingent claims. The perfect competition equilibrium without proper consumer information will therefore yield a $\left(\frac{P_i}{K_i}\right)^*$ ratio higher than would result with perfect information about the risk of nonpayment due to bankruptcy.⁷¹

There is certainly some validity to this argument. However, it would seem that capital markets themselves recognize the chance of bankruptcy very well. Where $\left(\frac{P_i}{K_i}\right)$ ratios become high enough to raise the probability of bankruptcy to more than *de minimis* proportions, the opportunity cost of capital probably begins to rise rather quickly, thus probably restricting the value of $\left(\frac{P_i}{K_i}\right)^*$ in a competitive market to some fairly reasonable level. There certainly does remain room for fraud and manipulation which should probably be a source of some public policy concern. This is especially true if management is not particularly sensitive to stockholder pressures and in mutual companies where the policyholders themselves are the source of the capital base. Therefore, the cartel pricing scheme certainly does tend to yield a $\left(\frac{P_i}{K_i}\right)^*$ ratio less than what would be achieved in a competitive market, but it is not at all clear that this increase in product

⁷⁰ An appendix discussing bankruptcy and consumer protection may be obtained upon request from the author.

⁷¹ And the efficient $\left(\frac{P_i}{K_i}\right)^*$ ratio in Figure 2(a) would then probably be finite without the necessity of having capital costs increase at an increasing rate with $\left(\frac{P_i}{K_i}\right)$.

quality is worth the price. A proposal is presented in the final section which should ensure that the risks associated with various $\left(\frac{P_i}{K_i}\right)$ ratios are properly taken into consideration by management and that consumers are protected from bankruptcy.

□ **Profit results for the property-liability insurance industry.** There has been a great deal of recent controversy over the profitability of the property-liability insurance industry.⁷² The analysis in the previous section indicated that since entry is easy we would expect the overall rate of return for the industry to be equal to the opportunity cost of capital. Although the Merritt Report (1911)⁷³ cast the question of profitability in terms of the ability to attract sufficient equity capital to provide service, this sensible notion of firm profitability, including all sources of income, was essentially lost from the regulatory process for over 50 years. The 1921 profit formula calling for a 5-percent return on premium volume gained acceptance as a just standard, even though no attempt had ever been made to show that this figure bore any relationship with rate of profit which would be sufficient to generate a rate of return just high enough to encourage entry and service. If profits fell below 5 percent of premiums, companies felt justified in asking that their rates be increased. Companies seeking to file deviated rates under prior approval laws were compelled to show that they would earn *at least* the 5 percent statutory profit figure.

It was not until the late 1960s that the profitability question again received detailed economic analysis. Under increasing pressures to return "high profits" to consumers (especially from some Senate Committees), the insurance industry commissioned a series of profitability studies by Arthur D. Little, Inc. (ADL).⁷⁴

The ADL studies reject the traditional profitability concept adopted by insurance regulators, at least for comparing profitability among industries. Instead, ADL conclude that it is the rate of return on capital that should be used in making such profitability comparisons. Included in return were all sources of income—underwriting profits, interest, dividends, and capital gains. The initial ADL study examined the profitability of the insurance industry for the period 1955 to 1965. In this study ADL chose to examine rates of return on total capitalization as the relevant measure of comparable profitability. For an industrial firm total capitalization includes stockholder's equity and surplus plus debt. For an insurance company the "equivalent" of total capitalization was taken to be total investable funds, which include equity capital and surplus *plus* unearned premium reserves. ADL argued that rate of return on total capital was the relevant rate of return figure for comparing different industries because this would give a measure of the relative "efficiencies" of capital employment among industries. The ADL studies also contain figures on rates of return on equity capital, but ADL

⁷² See National Association of Insurance Commissioners [16] for a discussion of the controversy and analysis of various measurement methodologies.

⁷³ Reported in U. S. Senate, Subcommittee on Antitrust and Monopoly [31], Part 5, p. 2791.

⁷⁴ Arthur D. Little, Inc. [1,2] and Irving Plotkin [24].

has continued to argue that those figures are not so relevant as are the rates of return on total capital.

The ADL studies use these figures in two ways. First ADL regressed the rates of return on total capitalization against the average intercompany temporal rate of return variance (a measure of risk) for a group of U. S. industries including the property-liability insurance industry (a sample of 43 stock companies). ADL found that the insurance industry point fell far below the regression line with an average rate of return on total investable funds of 4.4 percent and a risk measure of 10.89 (percent squared).⁷⁵ The study concludes that this indicates that the property-liability insurance industry is not earning a rate of return commensurate with that being earned in other industries with similar risks.

While the attempt to "adjust" rates of return by "risk" has great intuitive and theoretical appeal, the ADL reports, especially those following the initial report, have relied primarily on the straight comparison of rates of return between industries, using aggregate data from *Best's* to measure property-liability industry performance. For the period 1955 to 1967 ADL found that stock insurance companies earned an average rate of return on investable funds of 3.8 percent,⁷⁶ while the rate of return on total capital for all industrial companies was 9.8 percent during the same period.⁷⁷ ADL finds, in addition, that the property and liability insurance industry had an average rate of return lower than all 119 Standard and Poor's industrials.⁷⁸

ADL also presents calculations for rates of return on equity (net worth) for the property and liability insurance industry. After making what appear to be reasonable adjustments for peculiarities in insurance industry accounting practices, ADL concludes that the rate of return on net worth was 7.3 percent⁷⁹ for stock insurance companies for the period 1955 to 1968. (It was about 9 percent without the adjustments.) This compared with an average rate of return on net worth of 11.8 percent for Standard and Poor's industrials (1955 to 1967) and 9.1 percent for 102 gas and electric public utilities. ADL concludes that even if one chooses rate of return on net worth as a measure for comparison, the property and liability insurance industry does very poorly.

There have been a number of criticisms of the ADL study. Most have pointed to minor accounting problems and some have been quite ridiculous.⁸⁰ There appear to be two serious flaws in the ADL profitability studies, however. First, although a comparison of long-run rates of return on total capital (adjusted for risk) may indicate the relative efficiencies of capital usage in various industries, there is absolutely no reason to expect that the property and liability industry uses its total capital resources efficiently. On the contrary,

⁷⁵ Plotkin [24], p. 187.

⁷⁶ Arthur D. Little [2], p. 34.

⁷⁷ *Ibid.*, p. 37.

⁷⁸ *Ibid.*, p. 40.

⁷⁹ *Ibid.*, p. 45.

⁸⁰ Some of the criticisms have bordered on the ridiculous. See Norgaard and Schick [23]. They include as profits any increase in assets, whether or not such assets had attendant liabilities. With their measure a company could become very "profitable" by going deeply into debt.

the analysis in the previous section indicates that, although free entry will drive returns on equity capital to the competitive level, price regulation will result in an equilibrium position in which the ratio of premiums to equity capital is too small. Since increased premium reserves are directly related to premium volume, total investable funds (equity plus unearned premium reserves) will be *too large* relative to premium volume. We would expect *a priori* that the rate of return on total capital would be *lower* in the property and liability insurance industry than in an industry of equivalent risk characteristics, but which used its available capital efficiently. The results of the risk-return analysis of ADL are consistent, therefore, with the conclusion that there is excess capital in the property and liability insurance industry, and do not give us a good measure for answering the question of whether the industry earns "normal" profits or not. Other things being equal, the rate of return on net worth would be the relevant profitability measure to use if one were concerned with the industry's ability to attract capital. (There is no real debt financing.)

Is the solution to this problem then to recalculate the ADL risk-return results using the rate of return on net worth? The answer, I believe, is no. The measure of risk employed by the ADL study (temporal intercompany variance in rates of return) is simply not a good measure of the relative riskiness of a financial institution, whose assets are primarily liquid, and industrial corporations, whose assets are primarily fixed capital. An investor should certainly be willing to tolerate more variation in expected profits of an insurance firm which can easily liquidate its assets, than in a steel company which cannot. The comparison of the profitability of financial institutions with industrial corporations requires a much more sophisticated definition of risk.

One proper approach toward answering the profitability question properly might be to compare insurance firms with comparable financial institutions. It was suggested above that the insurance firm is best viewed as a levered investment trust. Table 7 presents First

TABLE 7
 RATES OF RETURN ON NET WORTH FOR FIRE AND CASUALTY COMPANIES
 AND INVESTMENT TRUSTS

YEAR	FIRE AND CASUALTY*	INVESTMENT TRUSTS**
1965	2.9	2.9
1964	4.8	2.9
1963	4.4	3.1
1962	3.9	2.7
1961	6.0	3.2
1960	5.8	3.7
1959	7.4	4.0
1958	5.6	4.9
1957	2.0	4.7
1956	2.2	4.7
1955	5.5	5.2
AVERAGE	4.5	3.4

*NET WORTH = POLICYHOLDERS' EQUITY (PRIMARILY AT MARKET VALUE)
 NET INCOME = UNDERWRITING PROFITS, INTEREST AND DIVIDENDS, AND
REALIZED CAPITAL GAINS

**NET WORTH = NET ASSETS AT MARKET VALUE
 NET INCOME = OPERATING PROFIT, INTEREST AND DIVIDENDS, NOT
 INCLUDING CAPITAL GAINS

National City Bank figures on rates of return on net worth for fire and casualty companies (property and liability) and for investment trust companies.⁸¹ These figures are illustrative and do not represent a rigorous "comparable" earnings study. The comparability of these figures suffers from one serious problem. The investment trust figures do not include realized capital gains, while the insurance figures do. A crude adjustment is possible using some of ADL's own figures. ADL calculates the ratio of operating income to policyholders' surplus to be 4.2 percent for the period 1955 to 1965 and the ratio of operating income plus *all* realized gains (including interest and dividends) at 5.4 percent.⁸² The elimination of all realized gains reduces the rate of return on policyholders' surplus by 22 percent. Let us assume that realized *capital* gains are about half of total realized gains, so that the elimination of realized capital gains would reduce the rate of return on net worth of insurance companies by 10 percent. Applying this figure to the First National City Bank figures, we get an "adjusted" rate of return on net worth of 4.1 percent for the insurance industry. This is somewhat higher than the 3.4 percent figure for investment trusts, which is exactly what we would expect.

The conclusion of this little exercise should be that, at least superficially, the returns earned by the property and liability insurance industry are not out of line with what we would expect when we compare industries which are realistically comparable. Most importantly, this discussion suggests that the issue of the profitability of the insurance industry is still an open question. Although the discussion above suggests that the rates of return observed for the property and liability insurance industry should approximate the competitive cost of capital, sound empirical verification is still an open area of research which should be pursued. Of special interest would be the more direct approach of estimating the cost of capital itself instead of relying on comparisons with other industries. A discussion of exactly how one might use the various approaches to the estimation of the cost of capital for insurance firms is outside the scope of this paper. However, it is an exercise that is certainly well worth pursuing.

But as the discussion in the previous section indicates, a rate of return analysis proves nothing about whether premium rates are too high or too low. Regulation and bureau rate making may still lead to prices which are higher than necessary and to excess industry capacity. Since entry into the industry continues to be vigorous, it appears that the rate of profit which bureaus and regulators have been able to achieve is sufficient to guarantee that overall industry capacity will be large enough to cover the demand for remunerative lines of insurance.

□ **The behavior of direct writers.** Finally an explanation of the behavior of direct writers must be attempted. Recall the following stylized facts regarding direct writers:

(1) Substantially lower underwriting costs than the industry as a whole;

⁸¹ First National City Bank of New York, as reprinted in Plotkin [24].

⁸² Plotkin [24], p. 192.

- (2) Deviations below bureau rates for many lines of insurance and for many classes of customers;
- (3) Moderate to high entry barriers into the "direct writers' " market;
- (4) Gradually increasing market share in most lines of insurance; and
- (5) Stricter underwriting policies than the industry as a whole.

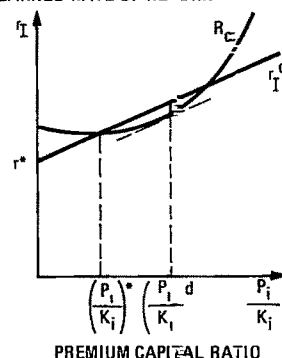
The surprising elements of this picture are (4) and (5). Since direct writers have much lower costs than agency companies, even after small deviations below bureau rates (to attract customers), it would seem that any risk class which is profitable for an agency company to write ($y_i \geq 0$) would be at least as profitable for the direct writers. They should therefore be willing to accept all risks that agency companies accept (perhaps even more) and rather *quickly* take over the entire market as the word that direct writers have lower rates gets around. The problem of lack of consumer information limiting the speed with which "the word gets around" has already been mentioned. But why do direct writers consciously refuse to write risks which agency firms in the voluntary market are quite willing to write?

Recall the model of long-run behavior of the firm discussed above. In particular let us concentrate on case (b) [Figure 2(b)]. Since the direct writers' market is characterized by a small number of firms and moderate to high barriers to entry, oligopolistic pricing behavior is possible. If successful oligopoly behavior is achieved there is no reason to expect that $\left(\frac{P_i}{K_i}\right)^*$ will result as an equilibrium position for

them. On the contrary, direct writers will seek a $\left(\frac{P_i}{K_i}\right)$ ratio where the difference between the earned rate of return and the cost of capital is maximized. (See Figure 3.) But the existing direct writers are certainly limited with regard to how quickly they can *individually* accumulate equity capital. As a result, in achieving the desired $\left(\frac{P_i}{K_i}\right)$ ratio they are limited with regard to the total premiums they can write in any period by the total equity capital which they have available. Adding premiums beyond this point increases their rate of return less than it increases the opportunity cost of capital.⁸³ Direct writers can achieve the profit-maximizing $\left(\frac{P_i}{K_i}\right)^d$ ratio in any period only by selling insurance to a portion of the profitable ($y_i \geq 0$) risks who demand it. Since they have to limit supply, they do best by choosing the best available risks in each class. Therefore, the strict underwriting policy is chosen. As equity capital is accumulated over time the direct writers would thus be expected to relax their underwriting policies, gaining a larger and larger share of the market.

In summary, the strategy of the direct writers in this scenario should be to lower their prices below bureau rates just enough to "differentiate" their products, and then fill up their premium "quota" with the best risks available. Even if entry barriers to direct writing

FIGURE 3
MAXIMIZATION OF DIFFERENCE
BETWEEN EARNED RATE OF
RETURN AND COST OF CAPITAL.
OPPORTUNITY COST OF CAPITAL,
EARNED RATE OF RETURN



⁸³ It is really stronger than this, since expected underwriting profits (I) will also decline as the poorer risks in each heterogeneous class are taken on.

TABLE 8

RETURN ON NET WORTH AFTER
TAXES AND POLICYHOLDER
DIVIDENDS OF 5 LARGEST
DIRECT WRITERS (1959-1968)

COMPANY	RATE OF RETURN
STATE FARM	13.1
ALLSTATE	20.9
LIBERTY MUTUAL	11.8
NATIONWIDE	7.4
GEICO	20.5
1200 PROPERTY- LIABILITY COMPANIES	8.8
SOURCE: U.S. SENATE, SUBCOMMITTEE ON ANTITRUST AND MONOPOLY [31], PART 17, p. 10201.	

are extremely high, they can exploit this "monopoly" power to only a limited extent, in a deregulated insurance market, because the competitive price for insurance written by agency firms will continue to provide a strong price ceiling.

There are two operating characteristics of insurance firms which should be consistent with the explanation of the behavior of direct writers. First, we would expect that direct writers will exhibit higher rates of return on net worth than the industry as a whole. Second, we should find that direct writers operate with a higher premium-equity capital ratio than the industry as a whole. A detailed analysis of these two indicators is still under way, but the following evidence is available at this time. Table 8 presents data on the average return on net worth after taxes and policyholder dividends of the five largest direct writers and the average rate of return for 1200 property-liability companies for the period 1959 to 1968. Four of the five direct writers exhibited rates of return substantially above the industry average, which is exactly what we expect from the hypothesis.

An examination of the relevant operating ratios for 1971 (a very good year for the insurance industry) is very revealing. Table 9

TABLE 9

PREMIUM-NET WORTH RATIOS AND RATES OF RETURN FOR FIVE DIRECT WRITERS AND THE TOTAL PROPERTY AND LIABILITY INDUSTRY, 1971

COMPANY	NET PREMIUMS NET WORTH	EARNED PREMIUMS NET WORTH	OPERATING INCOME NET WORTH	TOTAL RETURNS NET WORTH
ALLSTATE	2.02	1.87	12.8	21.8
STATE FARM	2.19	2.11	25.1	29.4
LIBERTY MUTUAL	3.52	3.51	19.8	25.0
GEICO	3.45	3.04	19.1	22.7
NATIONWIDE	2.09	2.06	17.7	27.8
TOTAL INDUSTRY	1.79	1.73	13.1	19.5
SOURCE: CALCULATED FROM NEW YORK STATE INSURANCE DEPARTMENT, STATISTICAL TABLES AND ANNUAL STATEMENT 1972.				

presents values for two measures of profitability—operating (insurance) income divided by net worth and total income (including interest, dividends, and capital gains) divided by net worth.

The model of the long-run behavior of direct writers indicates that direct writers should earn higher operating returns on net worth and operate at higher $\left(\frac{P}{K}\right)_i$ ratios than the industry as a whole. Table 9 presents additional evidence indicating that both predictions appear to be correct. All five direct writers operated at $\left(\frac{P}{K}\right)$ (where $\frac{P}{K}$ is calculated based on all of the companies' insurance operations) ratios larger than the average for the industry. Liberty Mutual and GEICO operated with $\left(\frac{P}{K}\right)$ almost twice the industry average. In addition, four of the five direct writers had substantially higher ratios of operating income to net worth than the average for the industry. The one company (Allstate) which had an operating return slightly below

the industry average also operated at a $\left(\frac{P}{K}\right)$ ratio closest to the industry average.

While the data presented in Tables 8 and 9 do not provide conclusive "proof" of the validity of the behavioral models presented here, the data are so consistent with the implications of these models that further consideration and attempts at empirical verification are certainly justified.

□ **Conclusions: why regulate insurance rates?** In light of the foregoing analysis we may conclude that the property-liability insurance industry under prior approval rate regulation has the following features:

- (1) A competitive market structure with a large number of firms and low levels of concentration;
- (2) Constant returns to scale in production;
- (3) Low to moderate barriers to entry for agency companies; moderate to high barriers to entry for direct writers;
- (4) A combination of rate making in concert and regulation which makes true price competition difficult;
- (5) An inefficient sales technique, probably costing consumers hundreds of millions of dollars per year, which is being eroded, but only slowly;
- (6) Supply shortages induced by the inability or the resistance of regulatory authorities to establish truly homogeneous risk classifications; and
- (7) Insurance premiums which are probably too high and effective capacity which is probably too large.

Faced with this picture, one wonders why insurance rates and rating classifications should be regulated at all. There are no natural monopoly characteristics which would indicate that open competition would be unstable and eventually lead to monopoly. Rather, the argument has been that rate making in concert through rating bureaus is a necessity to insure the public and the industry against "destructive" competition and large numbers of bankruptcies. There does not seem to be any reason why this industry should be more unstable than others as long as fraudulent practices are guarded against and proper consumer information is provided for. In many instances recommendations for deregulation of particular industries have rested on conclusions about performance after deregulation which are largely speculative. In the case of the property-liability insurance industry we do not have to rely on such speculation entirely, however. California has been operating under a "no-filing" regulatory statute since 1947. Rates and to some extent rating classes have essentially been regulated by the forces of free competition. While the insurance commissioner continues to have general supervisory power to guard against unfair prices or practices, his main job is to see that the free market works effectively as the regulator of prices.

The experience in California has been excellent; there have been neither mass bankruptcies nor price wars. While supply shortages continue to exist in some cases, this is primarily because of time lags inherent in adjusting rating classifications and collecting data to justify classification changes. The shortage problem appears to have

been less severe than in other states, however. Rating bureaus continue to exist, but their powers are strictly advisory. Insurance companies find that they have much more freedom to adjust quickly rates to changing loss situations. This has become very important as inflation has been rapidly increasing the cost of accident repairs. As pointed out before, there tends to be more pricing at off-bureau rates in California than in prior approval states.

The success of the California system has not gone unnoticed. A number of states have recently enacted California-type rating laws.⁸⁴ The impetus and research backing the move was provided by the staff of the New York Commission. Along with the NAIC, the New York Insurance Department has strongly encouraged the move to more price competition in states where the market is competitive. In addition, a new emphasis on providing consumers with price information for premium rates charged by different companies has been instituted in New York.⁸⁵ The Department realized that such information is an important component of any open competition system, especially in the context of the American Agency System, where price comparisons have historically been difficult.

On the basis of this analysis, the experience in California, and the experience in New York since the no-filing statute was enacted in 1970, the following general public policy recommendations are made with regard to the property-liability insurance industry.

(1) Prior approval rate regulation should be eliminated and replaced with a no-filing system allowing insurance prices to be determined competitively. State Insurance Departments should be retained to perform certain consumer protection functions to be outlined below. Insurance companies should file rate schedules with the insurance department as a source of general information. Companies should be free, however, to set any level of rates which they please.

(2) All anticompetitive aspects of rating bureaus should be eliminated. The rating bureaus should become strictly service organizations, collecting and processing loss and expense data for their customers. Any information provided to one company should be provided to all at appropriate fees.

(3) Flexibility in establishing truly homogeneous rating classifications should be encouraged. While a great deal of uniformity among companies is probably desirable with regard to the establishment of risk classifications so as to facilitate the collection of consistent loss data, no company should be forced to adhere to any established structure if it believes rate variability is justified. Given the availability of automatic data processing equipment, wise coding of all policyholder characteristics should make it possible for a rating bureau to provide loss experience to a customer insurance company for almost any conceivable category. Competition among rating bureaus may even be desirable.

(4) The Insurance Department should play a consumer information and consumer protection role. The greatest possible amount of price information should be put into the hands of consumers. Handbooks listing representative rates for all major underwriters should

⁸⁴ Among them are New York, Illinois, Florida, Ohio, Connecticut, and Indiana.

⁸⁵ See New York State Insurance Department [17, 18, 19].

be provided. Possible additional savings available from mutual companies paying dividends and participating stock companies should be brought to the attention of consumers. The insurance department should continue to audit company books and enforce the minimum capital requirement provisions of the state insurance laws. They should assist consumers by mediating complaints against insurance companies and by publicizing the names of companies which have consistently poor payoff policies.⁸⁶

(5) All insurance companies should be required to carry complete insurance against bankruptcy. Insurance rates should not be uniform for each company nor should a state insurance fund be made available to pay off for bankrupt companies. Neither of these schemes would give companies additional incentives to evaluate the premium-capital ratio they are carrying in terms of the true risk adjusted opportunity cost of capital. Rather, bankruptcy insurance rates should be geared to the insolvency risk of the companies themselves as determined by semiannual audits of their operations.

Companies operating with very high $\left(\frac{P_i}{K_i}\right)$ ratios should pay higher insolvency insurance rates than those operating with lower $\left(\frac{P_i}{K_i}\right)$ ratios. Companies which do not have to go to the capital markets for new funds will thus continue to face the risk consequences of their underwriting policies. It will help avoid attempts by risk-loving managers to run companies at very high $\left(\frac{P_i}{K_i}\right)$ ratios in the face of consumer ignorance and temporary impotence of capital market forces.

(6) Attempts should be made to speed up the transition from agency production of customers to direct writing wherever possible. The social costs of current laws forbidding agency companies from writing existing customers should be more thoroughly studied and an equitable scheme for phasing out independent agents devised. Trade associations of insurance agents should be strictly enjoined from taking any concerted action against a company which attempts to switch to direct writing region by region. Other barriers to entry of direct writers should be isolated and efforts made to lower them.

(7) Assigned risk pools will, by necessity, have to be continued as long as supply shortages continue to exist. Hopefully competition in rates and rating classification and extended consumer information will eventually cause the shortage problem to disappear. The practice of subsidizing high risk drivers in the assigned risk plans should be carefully reevaluated.

(8) Attempts by some states to go toward more price regulation rather than less should be vigorously discouraged. New Jersey has recently attempted to introduce public utility type rate-of-return rate

⁸⁶ The consumer protection function should also extend beyond rates to include standardization of basic policy forms and contract provisions. A wide array of policy provisions for basic types of customers would probably result in more confusion than expected benefits from "innovation" are worth. Companies or groups of companies actually creating "innovations" in types of coverage should be encouraged to institute them after scrutiny of the policy provisions by the regulatory authority.

regulation for insurance.⁸⁷ While the rate of return is certainly a much more meaningful profit figure than the rate of return on premiums, the attempt to determine a fair rate of return for the entire insurance industry is fraught with difficulties. How should the rate of return be calculated? What should be the base? Should we not calculate a fair rate of return for each company? Regulators attempting to apply public utility rate-of-return procedures to individual insurance firms or for the industry as a whole will be applying these techniques to an industry that has every single characteristic of historical regulatory failure. If there is no apparent reason to go this route, this should remain closed.

There are very few examples of deregulation in the insurance industry in the United States. Here is an example of a situation in which regulation has worked to the benefit of the consumer (except perhaps the direct writers) in the last 15 years. The industry designed to promote the orderly sale of fire insurance, which was a cartel protected by state regulation finally broke down. As the product changed, the sales technology changed, and the distribution attempts of regulators led to supply shortfalls. The short-run profitability for existing firms made almost everyone happy with the existing system. The availability of a model of a competitive market which really worked made the transition politically more appealing. The elimination of formal rate regulation should be encouraged in other states encouraged to follow their example.

References

1. ARTHUR D. LITTLE, INC. *Prices and Profits in the Property and Casualty Insurance Industry: Summary Report*. November 1967.
2. ———. *Studies of the Profitability, Industrial Structure, and Financial Condition of the Property and Liability Insurance Industry*. June 15, 1967.
3. BAIN, J. S. *Barriers to New Competition*. Cambridge: Harvard University Press, 1956.
4. *Best's Aggregates and Averages—Property and Liability Insurance*. New York: Alfred M. Best, Inc., published annually.
5. *Best's Insurance Reports—Property and Liability Insurance*. New York: Alfred M. Best, Inc., published annually.
6. *Best's Review*. New York: Alfred M. Best, Inc., published annually.
7. BRAINARD, C. H. AND DIRLAM, J. B. "Antitrust, Regulation, and the Insurance Industry: A Study in Polarity." *The Antitrust Bulletin* 14 (1969).
8. CRANE, F. G. *Automobile Insurance Rate Regulation: The Problem of Competition*. Ohio State University, Bureau of Business Research, 1968.
9. DIRLAM, J. AND STELZER, I. M. "The Insurance Industry and the Workability of Regulated Competition." *U.P.L.* 1 (1958).
10. HAMMOND, J. D. AND SCHILLING, N. "A Review Article—on Prices and Profits in the Property and Liability Insurance Industry." *Journal of Risk and Insurance* (March 1969).
11. HENSLEY, R. J. *Competition, Regulation and the Public Interest in Insurance*. Berkeley: Univ. of California Press, 1962.
12. KAHN, A. E. *The Economics of Regulation: Principles and Applications*. New York: John Wiley, 1970.
13. LONG, J. "Comments on the Plotkin Paper." *The Journal of Risk and Insurance* (June 1969).

14. LONG, J. D. AND GREGG, D. W., eds. *Property and Liability Insurance Handbook*. Homewood, Ill.: Richard D. Irwin, Inc., 1965.
15. NATIONAL ASSOCIATION OF INSURANCE COMMISSIONERS. "An Early Look at the Decision in the New Jersey Remand Case." May 8, 1972.
16. ———. "Measurement of Profitability and Treatment of Investment Income in Property and Liability Insurance." June 1970.
17. NEW YORK STATE INSURANCE DEPARTMENT. "The Automobile Assigned Risk Plan in New York State," Preliminary Report, 1973.
18. ———. "Competition among Auto Insurers in New York State." Preliminary Report, 1972 (as corrected by correspondence with Insurance Department).
19. ———. "Competition in Homeowners Insurance in New York State," Preliminary Report, 1973.
20. ———. "McCullough Report on Underwriting Profit." September 16, 1947.
21. ———. "The Public Interest Now in Property and Liability Insurance." January 7, 1969.
22. NEW YORK STATE LEGISLATURE. "Report of the Joint Committee of the Senate and Assembly of the State of New York, Appointed to Investigate Corrupt Practices in Connection with Legislation and the Affairs of Insurance Companies, Other Than Those Doing Life Insurance Business." February 1, 1911.
23. NORGAARD, R. AND SCHICK, G. "Profitability in the Property and Liability Insurance Industry." *The Journal of Risk and Insurance* (June 1968).
24. PLOTKIN, I. H. "Rates of Return in the Property and Liability Insurance Industry: A Comparative Analysis." *The Journal of Risk and Insurance* (June 1969).
25. POSNER, R. A. "Taxation by Regulation." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 22-50.
26. PRESIDENT'S NATIONAL ADVISORY PANEL ON INSURANCE IN RIOT-AFFECTED AREAS. *Meeting the Insurance Crises of Our Cities*. Washington, D.C.: U.S. Government Printing Office, 1968.
27. SCHERER, F. M. *Industrial Market Structure and Economic Performance*. New York: Rand McNally, 1970.
28. SHAPIRO, H. "Fire Insurance and the Inner City." The New York City-Rand Institute, R-703-NSF, February 1971.
29. SMALLWOOD, D. "Competition, Regulation and Product Quality in the Automobile Insurance Industry." Mimeo, April 28, 1972.
30. STELZER, I. M. "The Insurance Industry and the Antitrust Laws: A Decade of Experience." *The Insurance Law Journal* (March 1955).
31. UNITED STATES SENATE. Hearings Before the Subcommittee on Antitrust and Monopoly—The Insurance Industry—Parts 1-19. Washington, D.C.: U.S. Government Printing Office, 1959-1971.
32. UNITED STATES DEPARTMENT OF TRANSPORTATION. "Structural Trends and Conditions in the Automobile Insurance Industry." A Report of the Federal Trade Commission to the Department of Transportation, Washington, D.C.: U.S. Government Printing Office, April 1970.

The comparative statics properties of the theory of the regulated firm

David L. McNicol

Assistant Professor of Economics
University of Pennsylvania

This paper presents an analysis of the comparative statics properties of the theory of the regulated firm. It is assumed that the firm either maximizes profit, subject to regulatory constraint, or maximizes revenue, subject to regulatory constraint and a profit constraint. The question asked in each case is whether there are comparative statics properties which distinguish effective from ineffective regulation. Several properties which do so are identified. In particular, the firm's derived demands are found to have some very striking properties. It is shown that if the regulatory constraint is effective, then at given output the firm's demand for capital decreases while its demand for labor increases with increases in the wage rate. These peculiar properties also occur in some cases when the regulatory constraint is not binding if revenue is maximized. However, there prove to be properties which conclusively distinguish revenue maximization from profit maximization.

1. Introduction

■ Theoretical discussions of the model of the regulated firm proposed a decade ago by Averch and Johnson¹ have been largely concerned with clarifying the "A-J effect" and its implications for rate regulation. Relatively little effort has been devoted to developing the model as a tool for use in empirical research.² In particular, there has as yet been no systematic analysis of the model's comparative statics properties.

The well-known comparative statics properties of the model are:

- (1) The quantity of capital decreases with increases in the allowed rate of return;³
- (2) Output decreases with increases in the allowed rate of return if capital is not a regressive factor;⁴ and

David L. McNicol received the A.B. in economics from Harvard (1966), the S.M. in management from the Sloan School of Management (1968), and the Ph.D. in economics and finance from M.I.T. (1973). As Assistant Professor at the University of Pennsylvania, he is studying the economics of rate regulation. But it should be noted that the author considers research labels as specific as this "confining" and anything more accurate "vacuous."

A preliminary version of this paper was presented to the Industrial Organization Workshop of the University of Pennsylvania Department of Economics. The author benefited from comments made by members of the Workshop and especially from suggestions made by Jean Thomas Bernard, Noel Edelson, and Almarin Phillips. Thanks are also due an anonymous referee for expositional suggestions.

¹ In [2].

² See Baumol and Klevorick [7] for a survey of the literature.

³ In Takayama [23]. A proof of this proposition can also be found in Baumol and Klevorick [7].

⁴ Baumol and Klevorick [7].

- (3) The capital-labor ratio decreases with increases in the allowed rate of return if capital and labor are substitutes.⁵

Strictly analogous properties also hold for the usual model of unregulated monopoly. This similarity suggests the question: what gross differences are there between the comparative statics properties of regulated and unregulated monopoly?

An answer to this question, in addition to its general interest, could provide a formal basis for investigating whether regulation is “effective” in the sense that the allowed rate of return is less than the rate of return obtained by producing the monopoly output at cost.⁶ It would typically be assumed that the firm is best described as a profit-maximizing monopoly if regulation is ineffective. If this assumption were made and a property of the model that distinguishes it from unregulated monopoly proved to be inconsistent with observation, it presumably could be inferred that regulation was ineffective.

Several writers⁷ have questioned the assumption that regulated firms seek maximum profit, suggesting instead that they attempt to maximize revenue. This possibility is directly relevant to the issues of whether regulation is effective. There are plausible cases in which the regulatory constraint would be binding if profit were maximized, but would not be binding if revenue were maximized. Consequently, if observation indicated that the regulatory constraint was not binding, there could be a question of “Why not?” The answer would suggest a prescription for revision of the allowed rate of return. If profit is maximized and the regulatory constraint is not binding, it must be that the allowed rate of return is no less than the rate of return at the monopoly solution. A reduction in the allowed rate of return is clearly indicated in this case. However, it does not seem that a reduction in the allowed rate of return is called for if maximization of revenue leads to an output at which the regulatory constraint is ineffective.

This paper presents a comparative statics analysis of the model of the regulated firm. It is assumed in Section 2 that the firm maximizes profit subject to regulatory constraint and in Section 3 that instead the firm maximizes revenue subject to regulatory constraint and a profit constraint. The question asked in each section is whether there are comparative statics properties which distinguish effective from ineffective regulation. There prove to be several such properties. Section 4 asks whether comparative statics properties distinguish profit maximization from revenue maximization and identifies properties which do so.

■ Rate-of-return regulation proceeds on the basis of periodic comparisons of the rate of return actually earned by the firm with the “fair” or allowed rate of return set by the regulatory agency. If the firm has earned less than the allowed rate of return, it may request

2. The profit-maximization model

⁵ *Ibid.*

⁶ Several writers, most notably Stigler and Friedland [21], have argued that regulation in particular industries is ineffective in the sense that the allowed rate of return is as great as that which the firm could earn if it were unregulated. See also [8], [12], [16], and [18].

⁷ See [4], [7], and [13].

an increase in price with a reasonable prospect of success. If the firm has earned more than the allowed rate of return, the regulatory agency presumably will order the firm to reduce its price.

The regulatory constraint is assumed to be:

$$\frac{D(q, z)q - wL}{K} \leq s, \quad (1)$$

where K is capital (in value units), L is labor, w is the wage rate, s is the allowed rate of return, $P = D(q, z)$ is average revenue, $z = (z_1, \dots, z_n)$ is a vector of exogenous variables, and q is output. This statement of the constraint assumes away several problems, but it is adequate for an exploration of gross features of rate regulation. There is one matter of interpretation that should be recognized explicitly. Regulated firms are not required to satisfy the rate-of-return constraint instantaneously, so inequality (1) must be understood as applying to steady state equilibrium.

The profit-maximization problem assumed is

$$\max_{\{q, K, L\}} [\pi] = R(q, z) - [rK + wL], \quad (2)$$

subject to:

$$R(q, z) - [sK + wL] \leq 0, \quad (3)$$

$$q - \phi(K, L) \leq 0, \quad (4)$$

where r is the competitive cost of capital, $R(q, z)$ is total revenue and $\phi(K, L)$ is the production function. The production function is assumed to possess the following properties:

$\phi(K, L)$ is continuous and continuously twice differentiable;

$$\phi_K \equiv \frac{\partial \phi}{\partial K} > 0, \quad \phi_L \equiv \frac{\partial \phi}{\partial L} > 0, \quad K > 0, \quad L > 0;$$

$$\phi_{KK} \equiv \frac{\partial^2 \phi}{\partial K^2} < 0, \quad \phi_{LL} \equiv \frac{\partial^2 \phi}{\partial L^2} < 0, \quad K > 0, \quad L > 0;$$

$$\phi(0, L) = \phi(K, 0) = 0.$$

It is also assumed that:

$R(q, z)$ is continuous and continuously twice differentiable for $q \geq 0$;

$$\frac{\partial D}{\partial q} < 0, \quad \frac{\partial^2 R}{\partial q^2} < 0, \quad q \geq 0.$$

Finally, it is assumed that the markets for capital and labor are purely competitive and that the allowed rate of return is at least as great as the competitive cost of capital.

□ **The optimal solution.** There are two relevant types of optimal solution to the problem (2)–(4). First, if the allowed rate of return is at least as great as the rate of return at the monopoly solution, the monopoly solution is feasible and then, of course, the optimal solution. Second, if the allowed rate of return is less than the monopoly rate of return, the optimal output exceeds the monopoly output (if

capital is not a regressive factor) and is produced with more capital than is consistent with cost minimization. The question asked here is whether the comparative statics properties of these two solutions differ.

It is convenient to begin by developing a statement of the constrained case that parallels the usual model of unregulated monopoly. This statement is virtually essential for a clear comparison of the constrained and unconstrained cases. It, furthermore, considerably simplifies computation of comparative statics properties and provides a formulation of the model which is useful for econometric work.

The way the required statement can be developed is indicated by a simple observation about the problem. If (q^*, K^*, L^*) is an optimal solution, it must be that $rK^* + wL^*$ is at the smallest level permitted by the constraint. The problem can therefore be solved in two steps: (1) minimization of "cost" subject to the constraints; and (2) selection of the profit-maximizing level of output.⁸

A distinction will hereafter be maintained between "cost" and "expense." "Expense" ($\mathcal{E}$) will denote the actual outlays of the firm, while "cost" will have the usual meaning of minimal outlays for given production.

The expense minimization problem is:

$$\min_{\{K, L\}} [\mathcal{E}] = rK + wL; \quad (5)$$

subject to:

$$R(q, z) - [sK + wL] \leq 0, \quad (3)$$

$$L - \mathcal{L}(q, K) \geq 0, \quad (6)$$

$$K \geq 0, \quad L \geq 0, \quad (7)$$

where $\mathcal{L}(q, K)$ is the "labor requirements function" obtained by solving the production function. For given capital, $\mathcal{L}$ describes the quantity of labor required to produce output q and $\frac{\partial \mathcal{L}}{\partial q} \equiv \mathcal{L}_1 = 1/\phi_L$.

The graph of $\mathcal{L}$ for given output is an isoquant and $\frac{\partial \mathcal{L}}{\partial K} \equiv \mathcal{L}_2 = -\phi_K/\phi_L$. The assumptions that $\phi_L > 0$ and $\phi_{LL} < 0$ imply $\mathcal{L}_{11} > 0$. It is also assumed that $\mathcal{L}_{22} > 0$ (isoquants convex to the origin) and $\mathcal{L}_{12} = \mathcal{L}_{21} < 0$. Since $\mathcal{L}_{12} = -\frac{1}{\phi_L^2} \left(\frac{\partial \phi_L}{\partial K} \right)_q$, the latter assumption holds if and only if $\left(\frac{\partial \phi_L}{\partial K} \right)_q > 0$; i.e., the marginal productivity of labor increases with increases in capital along an isoquant. The importance of this assumption will be brought out shortly.

A statement of optimality conditions for an input combination to minimize expense is given in the Appendix. These conditions are not particularly useful. A clearer view of the problem and its solution is obtained by simply inspecting the feasible set.

It is useful, first, to specify the set of outputs which cannot be produced at cost. It is assumed: (1) that the rate of return decreases

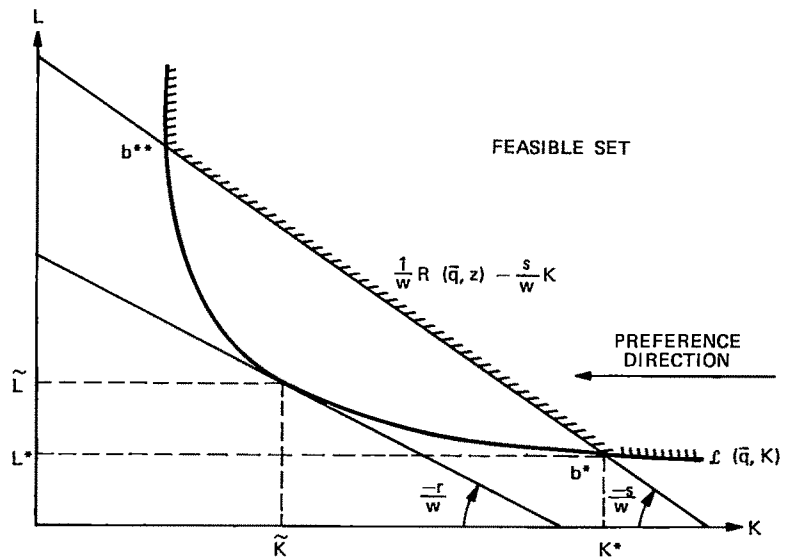
⁸ This approach follows Edelson [10] and to a lesser extent Klevorick [15].

Figure 1 is a graph of the feasible combinations of capital and labor for given output $0 < \bar{q} < q'$.¹⁰ The regulatory constraint requires that the quantities of capital and labor employed satisfy:

$$L \geq \frac{1}{w} R(\bar{q}, z) - \frac{s}{w} K. \quad (8)$$

The firm's input decision must also satisfy $L \geq \mathcal{L}(\bar{q}, K)$. The combination $(\bar{K}, \bar{L})$ required by cost minimization is infeasible since output $\bar{q}$ is assumed to be less than q' . The problem is to pick the feasible point with the smallest level of expense.

FIGURE 1
THE FEASIBLE SET FOR GIVEN OUTPUT $\bar{q} < q'$



Most of the feasible points can be immediately eliminated as candidates for the input combination which minimizes expense. It is clear that the optimal solution must lie on the boundary of the

⁹ It can be shown that this assumption is satisfied if the profit function is strictly concave for $q \geq 0$ and the capital-output ratio is constant or increases with output.

¹⁰ This figure (which is essentially the same as one used by Edelson [10]), contradicts Averch and Johnson's Figure 1. Averch and Johnson show the optimal solution as occurring at the point of tangency between an isoquant and a budget line with (in the notation used here) slope $-s/w$. This is apparently the basis of their statement (p. 1057) that for a fixed proportions production function the firm "... is constrained to the efficient expansion path." Both this statement and the diagram are incorrect. If the diagram were correct, $s + w\mathcal{L}_2 = 0$. But this implies $M\phi_K - s = 0$, since at the optimal solution $M = w/\phi_L$ (cf. Baumol and Klevorick [7]). The first-order conditions then require $M\phi_K - r = 0$, which is a contradiction since it is assumed that $s > r$.

feasible set.¹¹ Furthermore, given that isoquants are convex to the origin, expense increases as capital is reduced along the isoquant from its value at b^{**} . Expense also increases as capital is increased from its value at b^* . Consequently, the input combination that minimizes expense must lie on the segment of the boundary $b^{**}b^*$.

Let (K, L) be a point distant from (K^*, L^*) on this segment of the boundary of the feasible set. The constraint holds with strict equality at both (K^*, L^*) and (K, L) . Therefore:

$$\mathcal{E}^* = rK^* + wL^* = R(\bar{q}, z) - (s - r)K^*, \tag{9a}$$

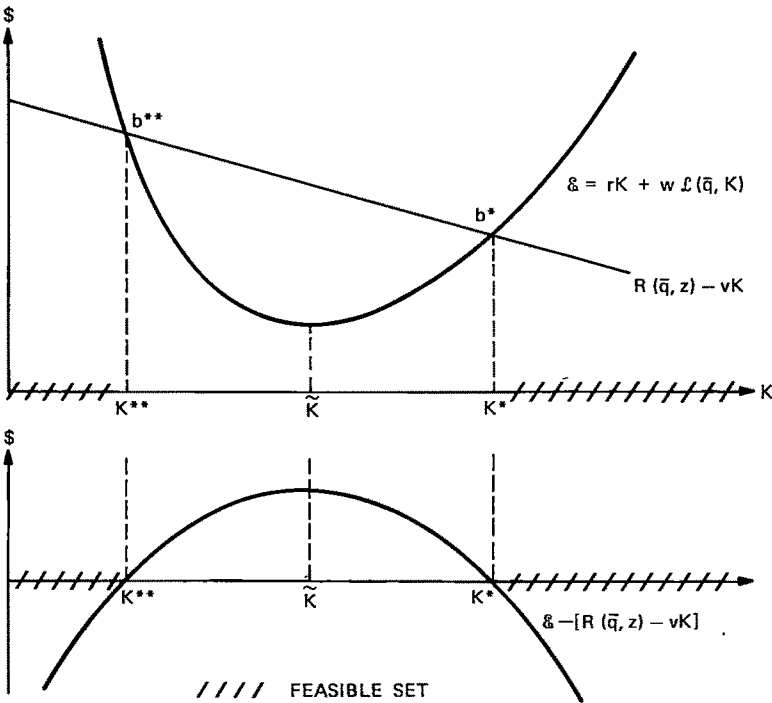
$$\mathcal{E} = rK + wL = R(\bar{q}, z) - (s - r)K; \tag{9b}$$

and the difference between expense at these two points is:

$$\mathcal{E}^* - \mathcal{E} = -v(K^* - K), \tag{9c}$$

where $v \equiv s - r$. If it is assumed that $v > 0$, then $\mathcal{E}^* < \mathcal{E}$ since $K^* > K$. Thus, expense is minimized at (K^*, L^*) . As is shown in Figure 2, this is the case because $R(\bar{q}, z) - vK$ ("required expense") decreases with increases in capital.¹²

FIGURE 2
EXPENSE MINIMIZATION



¹¹ Bailey [3] shows that a firm which attempts to maximize either profit or revenue subject to a rate-of-return constraint will never waste inputs. In an extended discussion of waste, Bailey also shows that the firm may waste input in some circumstances if it is subject to other forms of regulation. See also, Edelson [10] and Zajac [26, 27].

¹² There has been some discussion, beginning with Takayama [23], and continuing with Pressman and Carol [19, 20] and El-Hodiri and Takayama [11], about what assumptions are required to prove the "A-J effect." As the preceeding argument brings out, four assumptions are required. It is first necessary to assume

The fact that profit maximization requires expense minimization can be used to obtain the firm's derived demands and the total expense function. It is assumed that isoquants are convex to the origin, so for any output $\bar{q} < q'$ there are (cf. Figure 1) two values of capital for which

$$R(\bar{q}, z) - [sK + w\mathcal{E}(\bar{q}, K)] = 0, \quad 0 < \bar{q} < q'. \quad (10a)$$

Assuming that $\nu > 0$, it is the larger of these two values which is relevant as it yields the smaller expense. The slope of the isoquant ($\mathcal{E}_2 = -\phi_K/\phi_L$) is greater than $-s/w$ at the input combination which minimizes the expense of producing an output $q < q'$. It follows that

$$\frac{\partial \psi}{\partial K} = s + w\mathcal{E}_2^* > 0, \quad 0 < q < q', \quad (10b)$$

where $\mathcal{E}_2^*$ is the slope of the isoquant evaluated for the value of capital which minimizes expense and $\psi \equiv R(q, z) - [sK + w\mathcal{E}(q, K)]$. Equation (10a) can therefore be solved for the expense minimizing value of capital. Note that at output q' , (10a) is satisfied only for the value of capital which minimizes cost. The expense minimizing quantity of labor is obtained by substituting the relationship for capital into the labor requirements function. These relationships are denoted:

$$K = E^1(q, s, w, z), \quad \nu > 0, \quad 0 < q \leq q', \quad (11a)$$

$$L = E^2(q, s, w, z), \quad \nu > 0, \quad 0 < q \leq q'. \quad (11b)$$

Equations (11a) and (11b), which are simply a parametric representation of a portion of the boundary of the feasible set, can be interpreted as the expense minimizing derived demands for capital and labor. Substitution of the derived demand equations into the definition of $\mathcal{E}$ gives the total expense function.

Figure 3 presents a graph of the marginal expense function. It is assumed for convenience that marginal cost is a constant C' . Marginal expense is denoted $\mathcal{E}'$ and marginal revenue is denoted $M(q, z)$. The marginal expenses of increasing output by increasing only labor (evaluated for the expense minimizing value of capital) is $w\mathcal{E}_1^*$. Three important properties of marginal expense appear in this figure:

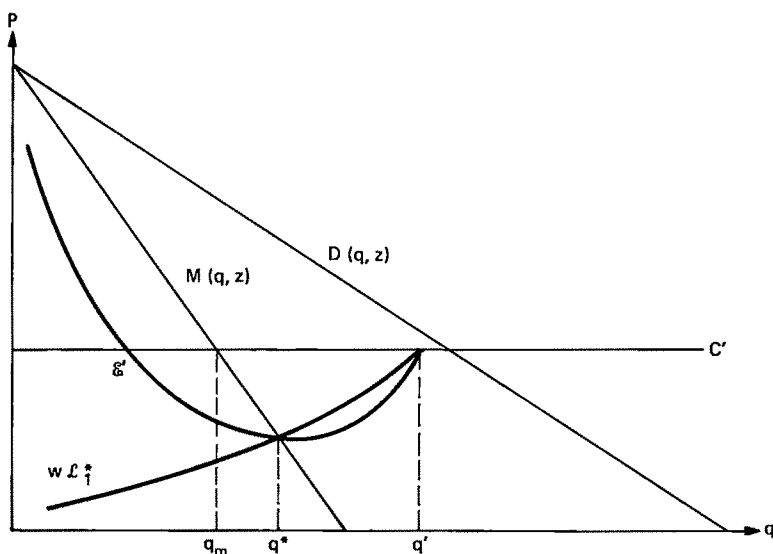
- (1) At q' , $\mathcal{E}' = C'$;
- (2) $w\mathcal{E}_1^* < C'$, $\nu > 0$, $0 < q < q'$; and
- (3) Marginal expense is a weighted average of marginal revenue and $w\mathcal{E}_1^*$; i.e.,

$$\mathcal{E}' = \theta^* M + (1 - \theta^*) w\mathcal{E}_1^*, \quad 0 \leq \theta^* < 1, \quad \nu > 0, \quad 0 < q \leq q'. \quad (12)$$

These properties are demonstrated in the Appendix. The proof that $w\mathcal{E}_1^* < C'$, for $\nu > 0$, and $0 < q < q'$, assumes that the marginal

that the output in question can not be produced at cost, and second, that there exists a feasible solution. The existence of a feasible solution can be guaranteed by making appropriate continuity assumptions, but these are not required as such. Third, it is necessary to assume that the isoquants are strictly convex to the origin. Finally, it must be assumed that $s > r$, since if $s = r$ the same level of expense can be achieved by either under- or overcapitalization.

FIGURE 3

MARGINAL EXPENSES, $w\mathcal{L}_1^*$, AVERAGE COST AND AVERAGE REVENUE

productivity of labor increases with increases in capital along an isoquant.

The second part of the problem is that of selecting the output which maximizes profit. It is assumed that the allowed rate of return exceeds the competitive cost of capital and it is relevant to assume further that the allowed rate of return is less than the rate of return at the monopoly solution; i.e., in terms of Figure 3, that the monopoly output q_m is less than q' .

It is reasonable to assume that the profit obtained if cost is minimized decreases monotonically for outputs $q \geq q_m$. Then, since in particular $q' > q_m$, attention can be limited to outputs $q \leq q'$, and the profit maximization problem is:

$$\max_{\{q \leq q'\}} [\pi] = R(q, z) - \mathcal{E}(q, s, w, z). \quad (13)$$

Using equation (12), the first-order condition for $0 \leq q^* \leq q'$ to maximize profit is

$$M^* - [\theta^* M^* + (1 - \theta^*) w\mathcal{L}_1^*] = 0; \quad (14a)$$

$$(1 - \theta^*)(M^* - w\mathcal{L}_1^*) = 0. \quad (14b)$$

It is shown in the Appendix that if $\nu > 0$, $0 \leq \theta^* < 1$ for $0 \leq q \leq q'$. Therefore, (14b) requires¹³

$$M^* - w\mathcal{L}_1^* = 0. \quad (14c)$$

It is clear from Figure 3 that if the unconstrained monopoly problem has a unique solution $q_m > 0$, and $w\mathcal{L}_1^* < C'$ for $0 < q < q'$ while

¹³ Substitution of $w\mathcal{L}_1 = w/\phi_L = M$ into the first-order condition for expense minimization given in the Appendix yields:

$$(M\phi_K - r) - \theta^*(M\phi_K - s) = 0.$$

This and (14c) restated as $M\phi_L - w = 0$ (and $\theta^* \geq 0$) are the first-order conditions usually stated.

$w\mathcal{L}_1^* = C'$ at q' , then there is an output q^* , $q_m < q^* < q'$, for which (14c) is satisfied.¹⁴ It is assumed that there is only one such output. The second-order condition is:

$$M'^* - \left[w\mathcal{L}_{11}^* + w\mathcal{L}_{12}^* \frac{\partial E^1}{\partial q} \right] < 0, \quad (15a)$$

where $M' = \frac{\partial^2 R}{\partial q^2}$. Differentiation of (10a) gives $\frac{\partial E^1}{\partial q^*} = \frac{M^* - w\mathcal{L}_1^*}{s + w\mathcal{L}_2^*} = 0$, so the second-order condition reduces to:

$$M'^* - w\mathcal{L}_{11}^* < 0. \quad (15b)$$

It is assumed that $M' < 0$ for all $q \geq 0$, and $w\mathcal{L}_{11} > 0$ for $q > 0$, $K > 0$, so (15b) is satisfied.

The conclusion that q^* is greater than the monopoly output rests on the assumption that $w\mathcal{L}_{12} < 0$; i.e., that the marginal productivity of labor increases with increases in capital along an isoquant. If $w\mathcal{L}_{12} > 0$, then $w\mathcal{L}_1^* > C'$ for $0 < q < q'$, and the optimal output q^* is less than q_m .

There is one other characteristic of the profit-maximization problem that is worth noting. It is assumed that q^* is unique and that marginal revenue decreases monotonically with output. Furthermore, at q' , $M - w\mathcal{L}_1^* = M - C'$ and, assuming that the monopoly solution is unique, $M - C' < 0$ at q' . Therefore, $M - w\mathcal{L}_1^* \geq 0$ for $q \leq q^*$ and $M - w\mathcal{L}_1^* < 0$ for $q^* < q \leq q'$. Since $s + w\mathcal{L}_2^* > 0$, it follows that $\frac{\partial E^1}{\partial q} = \frac{M - w\mathcal{L}_1^*}{s + w\mathcal{L}_2^*} < 0$ for $q^* < q \leq q'$. That is, the value

of capital *increases* as output is *reduced* from q' to q^* . And $\frac{\partial E^1}{\partial q} \geq 0$ for $q \leq q^*$, which is to say that the optimal solution occurs at the point on the boundary of the feasible set with the largest value of capital.¹⁵

□ **Comparative statics.** The discussion now turns to the question of what differences there are between the comparative statics properties of the constrained and the unconstrained solutions to the profit-maximization problem.

The statement of the model that has been developed permits two sorts of comparisons to be made. First, the properties of the expense minimizing derived demands can be compared with those of the cost minimizing derived demands. Two distinct ranges of output appear in Figure 3: (1) for outputs $q \geq q'$, cost is minimized; and (2) for

¹⁴ Westfield [24] presents an informal argument that $M^* > 0$. Baumol and Klevorick [7] state that it is necessary to assume that the optimal solution occurs at an output for which marginal revenue is positive. Bailey [3] and Dayan [9] show that if there then is an optimal solution to the problem, $M^* > 0$. However, it is unnecessary to assume directly that the constrained problem possesses a non-trivial solution or that marginal revenue is positive at the optimal solution. Baumol and Klevorick show that the output of the regulated firm exceeds the monopoly output if and only if:

$$\phi_K \phi_{LL} - \phi_L \phi_{KL} < 0.$$

This is equivalent to the assumption that $w\mathcal{L}_{12} < 0$.

¹⁵ This fact has been pointed out by Zajac [26] and Baumol and Klevorick [7].

outputs $q < q'$, expense is minimized. Which of these two intervals contains the optimal solution depends on the magnitudes of q_m and q' . If $q_m > q'$ (i.e., the allowed rate of return is greater than the rate of return at the monopoly solution), profit is maximized by producing the monopoly output at cost. If $q_m < q'$, the constraint is binding at the optimal output q^* , which is produced at minimal expense. Differences between the properties of the cost and expense minimizing derived demands would distinguish these two possibilities. Second, the signs of total derivatives of the optimal solutions can be compared. Both of these comparisons indicate properties that distinguish effective from ineffective regulation, but the first proves to be the more interesting.

TABLE 1
PARTIAL DERIVATIVES OF DERIVED DEMAND AND TOTAL DERIVATIVES OF THE OPTIMAL SOLUTION, PROFIT MAXIMIZATION WITH EFFECTIVE REGULATION

DERIVED DEMAND $v > 0, 0 < q < q'$	TOTAL DERIVATIVES $0 < q < q'$
$\frac{\partial E^1}{\partial q} = \frac{M - wL_1^*}{s + wL_2^*}$ (16a)	$\frac{dq^*}{ds} = \mu^* \frac{\partial E^{1*}}{\partial s}$ (18a)
$\frac{\partial E^1}{\partial r} = 0$ (16b)	$\frac{dq^*}{dr} = 0$ (18b)
$\frac{\partial E^1}{\partial s} = -\frac{K}{s + wL_2^*}$ (16c)	$\frac{dq^*}{dw} = \mu^* \frac{\partial E^{1*}}{\partial z_1}$ (18c)
$\frac{\partial E^1}{\partial w} = -\frac{L}{s + wL_2^*}$ (16d)	$\frac{dq^*}{dz_1} = -\mu^* \left[\frac{\partial M^* / \partial z_1}{wL_{12}^*} - \frac{\partial E^{1*}}{\partial z_1} \right]$ (18d)
$\frac{\partial E^1}{\partial z_1} = \frac{\partial R / \partial z_1}{s + wL_2^*}$ (16e)	$\frac{dE^{1*}}{d\alpha} = \frac{\partial E^{1*}}{\partial \alpha}$ (19) $\alpha \in (s, r, w, z)$
$\frac{\partial E^2}{\partial q} = L_1^* + L_2^* \frac{M - wL_1^*}{s + wL_2^*}$ (17a)	$\frac{dE^{2*}}{d\alpha} = -\epsilon^* L_2^* \frac{\partial E^{1*}}{\partial \alpha}$ (20a) $\alpha \in (s, r, w)$
$\frac{\partial E^2}{\partial r} = 0$ (17b)	$\frac{dE^{2*}}{dz_1} = L_1^* \frac{dq^*}{dz_1} + L_2^* \frac{\partial E^{1*}}{\partial z_1}$ (20b)
$\frac{\partial E^2}{\partial s} = -L_2^* \frac{K}{s + wL_2^*}$ (17c)	$\mu^* = \frac{wL_{12}^*}{M^* - wL_{11}^*}$
$\frac{\partial E^2}{\partial w} = -L_2^* \frac{L}{s + wL_2^*}$ (17d)	$\epsilon^* = \frac{wL_{12}^*}{\phi_K^* (M^* - wL_{11}^*)} - 1$
$\frac{\partial E^2}{\partial z_1} = L_2^* \frac{\partial R / \partial z_1}{s + wL_2^*}$ (17e)	

Table 1 presents the partial derivatives of the expense minimizing derived demands and the total derivatives of the optimal solution for the case in which regulation is effective. The partial derivatives of the expense minimizing derived demand for capital are obtained by differentiating (10a). The corresponding partial derivatives of the derived demand from labor are then obtained from $\mathcal{L}(q, K)$. The total

derivatives of the optimal output with respect to each of the parameters are obtained by differentiating the first-order condition (14c). The total derivatives of capital and labor with respect to parameter α are:

$$\frac{dE^{1*}}{d\alpha} = \frac{\partial E^{1*}}{\partial q} \frac{dq^*}{d\alpha} + \frac{\partial E^{1*}}{\partial \alpha}, \quad (21a)$$

$$\frac{dE^{*2}}{d\alpha} = \mathcal{L}_1^* \frac{dq^*}{d\alpha} + \mathcal{L}_2^* \frac{dE^{1*}}{d\alpha}. \quad (21b)$$

But $\frac{\partial E^{1*}}{\partial q} = 0$, so $\frac{dE^{1*}}{d\alpha} = \frac{\partial E^{1*}}{\partial \alpha}$. Substitution of the formulas for $\frac{dq^*}{d\alpha}$ and $\frac{dE^{1*}}{d\alpha}$ into (21b) yields (20a) on simplification.

The sign restrictions obtained¹⁶ are summarized in Table 2 along with the corresponding properties of the monopoly solution.¹⁷ To

TABLE 2
COMPARATIVE STATICS PROPERTIES OF THE PROFIT MAXIMIZATION MODEL

(A) COMPARATIVE STATICS OF DERIVED DEMAND										
CASES	$\frac{\partial K}{\partial ()}$					$\frac{\partial L}{\partial ()}$				
	q	s	r	w	z_1	q	s	r	w	z_1
OVER-CAPITALIZATION $q < q'$	$\pm$	—	0	—	$\text{sign } \frac{\partial D}{\partial z_1}$	?	+	0	+	$-\text{sign } \frac{\partial D}{\partial z_1}$
COST MINIMIZATION $q \geq q'$	+	0	—	+	0	+	0	+	—	0

(B) TOTAL VARIATIONS IN THE OPTIMAL SOLUTION												
CASES	$dq^*/d()$				$dK^*/d()$				$dL^*/d()$			
	s	r	w	z_1	s	r	w	z_1	s	r	w	z_1
EFFECTIVE REGULATION $q_m < q'$	—	0	—	$\text{sign } \frac{\partial D}{\partial z_1}$	—	0	—	$\text{sign } \frac{\partial D}{\partial z_1}$	?	0	?	?
INEFFECTIVE REGULATION $q_m \geq q'$	0	—	—	$\text{sign } \frac{\partial D}{\partial z_1}$	0	—	?	$\text{sign } \frac{\partial D}{\partial z_1}$	0	?	—	$\text{sign } \frac{\partial D}{\partial z_1}$

verify the restrictions shown for the case in which regulation is effective, recall that $s + w\mathcal{L}_2^* > 0$ and that $M'^* - w\mathcal{L}_{11}^* < 0$. These two facts, along with the assumptions that $\mathcal{L}_2 < 0$ and $\mathcal{L}_{12} < 0$ give im-

¹⁶ Note that $L \frac{\partial k}{\partial ()} = \frac{\partial E^1}{\partial ()} - k \frac{\partial E^2}{\partial ()}$, where $k = K/L$. The restrictions established then require $\frac{\partial k}{\partial s} < 0$, $\frac{\partial k}{\partial w} > 0$, and $\text{sign } \frac{\partial k}{\partial z_1} = \text{sign } \frac{\partial R}{\partial z_1}$.

¹⁷ Denoting the cost-minimizing derived demands by $K = F^1(q, r, w)$ and $L = F^2(q, r, w)$, it can be shown that $\frac{\partial F^1}{\partial q} > 0$ if $\mathcal{L}_{12} < 0$ and $\frac{\partial F^2}{\partial q} > 0$ if $\mathcal{L}_{12} < 0$ and $\mathcal{L}_{22} > 0$.

mediately the sign restrictions shown for the partial and total derivatives with respect to s , w , and z . It was noted earlier than $\frac{\partial E^1}{\partial q} < 0$

for $q^* < q \leq q'$. Then (17a) implies $\frac{\partial E^2}{\partial q} > 0$ for $q^* < q \leq q'$, but it is not clear that this is the case of $q \leq q^*$.

The most striking properties that appear in Table 2 are the "perverse" behaviors of the expense minimizing derived demands with respect to the wage rate. The expense minimizing derived demand for capital *decreases* while the expense minimizing demand for labor *increases* with *increases* in the wage rate. This is a very peculiar result, but it is not difficult to grasp factors which lie behind it. An increase in the wage rate increases the expense associated with any input combination $K > 0$, $L > 0$. In terms of Figure 1, this is represented by a shift in the boundary of the feasible set towards the left. Consequently, the firm can become *less inefficient* and, since it is assumed that expense is minimized, it will do so. This requires decreasing the quantity of capital employed and correspondingly increasing the quantity of labor employed.

The partial derivatives of the expense minimizing derived demands for capital and labor with respect to the allowed rate of return have the same sign as the corresponding partial derivatives of the cost minimizing derived demands with respect to the competitive cost of capital. The logic behind these two cases is, however, quite different. The response of the expense minimizing derived demands must, again, be understood in terms of changes in the feasible set. An increase in the allowed rate of return, like an increase in the wage rate, shifts the boundary of the feasible set shown in Figure 1 to the left, which leads the firm to use less capital and more labor.

The expense minimizing derived demands have two other unusual properties. First, for most technologies, cost minimization would require that capital be a strictly increasing function of output while the sign of $\frac{\partial E^1}{\partial q}$ is ambiguous. Second, the exogenous variables of the average revenue function appear in the expense minimizing derived demand functions. This simply reflects the fact that the quantities of capital and labor which must be employed to satisfy the regulatory constraint depend on the level of revenue. Note that an increase (decrease) in revenue at given output requires the regulated firm to become less (more) efficient. This in turn requires an increase (decrease) in the capital stock and a decrease (increase) in the quantity of labor employed.

It is important to recognize that the results obtained for the derived demands hold not only for the optimal output but for *any* output less than q' . They are as much descriptions of how the boundary of the feasible set varies as descriptions of the behavior of the firm. In the present static context, these restrictions can fail *only* if: (1) the firm does not minimize expense; or (2) the regulatory constraint is not binding so the firm produces its output at cost.

The unusual properties of the derived demands do not seem to necessarily appear in the total derivatives of the optimal solution. It can be shown that $\epsilon^* < 0$ and hence $\frac{dE^{*2}}{ds} > 0$ and $\frac{dE^{*2}}{dw} > 0$ if and

only if¹⁸:

$$\frac{\partial^2 R^*}{\partial K^* \partial L^*} = M'^* \phi_K^* \phi_L^* + M^* \phi_{KL}^* < 0. \quad (22)$$

The restriction that $\frac{dE^{*2}}{dw}$ be positive would be useful, as the corresponding derivative of the monopoly solution is negative. However, Inequality (22) is not necessarily satisfied.

Table 2(B) does nonetheless contain one useful set of restrictions. If regulation is effective, then the profit-maximizing values of output, capital, and labor are independent of the competitive rate of return.¹⁹ This, of course, is not true of the monopoly solution.

The restrictions that appear in Table 2 are primarily useful as a means of extracting conclusions from an econometric model of a regulated firm or sample of regulated firms. A literal (and linear) translation of the profit-maximization model into econometric terms is:²⁰

$$q = \alpha_0 + \alpha_1 s + \alpha_2 r + \alpha_3 w + \alpha_4 z + e_1, \quad (23)$$

$$K = \beta_0 + \beta_1 s + \beta_2 r + \beta_3 w + \beta_4 z + \beta_5 q + e_2, \quad (24)$$

$$L = \gamma_0 + \gamma_1 s + \gamma_2 r + \gamma_3 w + \gamma_4 z + \gamma_5 q + e_3, \quad (25)$$

where e_1 , e_2 , and e_3 are error terms. If regulation were known to be effective, solution of the first-order condition (14c) would (assuming the linear form) be translated to:

$$q = \alpha_0 + \alpha_1 s + \alpha_3 w + \alpha_4 z + e_1. \quad (23a)$$

Similarly, if regulation were known to be ineffective, solution of the usual first-order condition would suggest:

$$q = \alpha_0 + \alpha_2 r + \alpha_3 w + \alpha_4 z + e_1. \quad (23b)$$

These two possibilities are combined in (23). It would be inferred that regulation was effective if $\hat{\alpha}_1 < 0$ and $\hat{\alpha}_2 = 0$, where carets denote estimated values. Conversely, if $\hat{\alpha}_1 = 0$ and $\hat{\alpha}_2 < 0$, it would be inferred that regulation was ineffective. (In both cases, it would be expected that $\hat{\alpha}_3 < 0$.) Equations (24) and (25) are of course simply

¹⁸ Differentiate $\mathcal{L}_1 = 1/\phi_L$ and $\mathcal{L}_2 = -\phi_K/\phi_L$ and use $w\mathcal{L}_1^* = M^*$ to obtain:

$$w\mathcal{L}_{12}^* = -M^* \frac{\phi_{LK}}{\phi_L} - \phi_K w\mathcal{L}_{11}^*.$$

Substitution of this expression into the definition of ϵ^* leads directly to inequality (22). Baumol and Klevorick [7] show that if this inequality is reversed, then $\frac{d(K^*/L^*)}{ds} < 0$. The assumption that $\frac{\partial^2 R^*}{\partial K^* \partial L^*} > 0$ plays a crucial role in Sheshinski [21]. Sheshinski shows that if this condition holds, then it is efficient to set the allowed rate of return at less than the rate of return at the monopoly solution.

¹⁹ Bailey [3] obtains $\frac{dq^*}{ds} = \frac{dE^{1*}}{ds} = 0$ by a somewhat different method. She also obtains $\frac{dE^{1*}}{dw} < 0$ and $\frac{dq^*}{dw} < 0$.

²⁰ Since $\frac{\partial E^{1*}}{\partial q} = 0$, any linear specification of the model would probably omit output from the list of variables in the equation for the derived demand for capital. Note that each of the equations in the system is identifiable if the error terms are contemporaneously uncorrelated.

a particular specification of the derived demand equations. It would be inferred that expense is minimized if it were found that:

$$\begin{aligned}\hat{\beta}_1 < 0 \quad \hat{\beta}_2 = 0 \quad \hat{\beta}_3 < 0, \\ \hat{\gamma}_1 < 0 \quad \hat{\gamma}_2 = 0 \quad \hat{\gamma}_3 > 0.\end{aligned}$$

Conversely, if it were found that:

$$\begin{aligned}\hat{\beta}_1 = 0 \quad \hat{\beta}_2 < 0 \quad \hat{\beta}_3 < 0, \\ \hat{\gamma}_1 = 0 \quad \hat{\gamma}_2 < 0 \quad \hat{\gamma}_3 < 0,\end{aligned}$$

it would be inferred that cost is minimized and hence that regulation is ineffective.

It would, of course, be necessary to adapt the theoretical model to any particular case. For example, the model as it stands is static, and it would probably not be reasonable to assume that observations were drawn from positions of steady state equilibrium. However, the procedure for drawing inferences from any estimated form of the model is basically as outlined.²¹

■ Some would object to such a test of the effectiveness of rate regulation on the grounds that regulated firms pursue maximum revenue rather than maximum profit. As noted earlier, there are plausible cases in which the constraint is binding if profit is maximized but is not binding if revenue is maximized.

This section considers the problem of revenue maximization subject to regulatory constraint. The question asked is, again, whether comparative statics properties distinguish effective from ineffective regulation. The task of identifying properties that distinguish profit maximization from revenue maximization, which proves to be simple, is taken up in the following section.

The revenue-maximization problem, assuming that the firm is subject to a constraint imposed by owners and/or creditors that at least some "satisfactory" rate of return $\bar{r}$ is earned, is:

$$\max_{\{q, K, L\}} [R] = D(q, z)q, \quad (26)$$

subject to:

$$R(q, z) - [\bar{r}K + wL] \geq 0; \quad (27)$$

$$R(q, z) - [sK + wL] \leq 0; \quad (3)$$

$$q - \phi(K, L) \leq 0. \quad (4)$$

Inequality (27) will be referred to as the profit constraint. It is assumed that $r \leq \bar{r} < s$ (as if not, the only feasible solution is $q = K = L = 0$).

Previous discussions of this problem have overlooked the fact that apart from one special case, the revenue-maximization problem possesses infinitely many optimal solutions for capital and labor.²²

²¹ See MacAvoy [17] for an example of the use of comparative statics restrictions in the study of the behavior of regulated firms.

²² In his original development of the model, Baumol [6] assumed that the firm minimizes cost. This eliminates the problem of infinitely many solutions. However, as was suggested independently and almost simultaneously by Bailey and Malone [4] and Kafoglis and Bushnell [14], a firm which maximizes revenue subject to a satisfactory rate of return may have an incentive *not* to minimize cost.

3. Revenue maximization

This situation arises because unlike the profit-maximization problem, the revenue-maximization problem does not embody any criterion for distinguishing among feasible combinations of capital and labor which yield the same output.

The most obvious criterion, and the one assumed, is total expense. That is, it is assumed that at each output, the firm selects the feasible combination (K, L) which yields the smallest total outlays.²³

□ **The optimal solution.** Revenue is at an unconstrained maximum at the output $\hat{q}$ (assumed to be finite) for which $M(\hat{q}, z) = 0$. The optimal output can then be simply described as the largest feasible output less than or equal to $\hat{q}$. If $\hat{q} < q'$, it is optimal to produce $\hat{q}$ at minimum expense, which requires overcapitalization. If $\hat{q} > q'$, it may be feasible to produce $\hat{q}$ at cost, which is then optimal. Neither of these cases²⁴ requires any further elaboration. The remaining possibility is that while $\hat{q} > q'$ it is not feasible to produce $\hat{q}$ at cost. In this case there is an incentive for undercapitalization, a possibility first²⁵ pointed out by Bailey and Malone.²⁶

The key to the last of these possibilities is the recognition that for outputs less than q_c (i.e., the output for which average revenue equals average cost), a larger rate of return can be obtained by undercapitalization than by producing the output at cost. This fact is shown graphically in Figure 4. It can be demonstrated as follows.²⁷ The rate of return at output $\bar{q} > 0$ is:

$$\rho \equiv \frac{R(\bar{q}, z) - w\mathcal{L}(\bar{q}, K)}{K}. \quad (28)$$

Let $\bar{K}$ be the value of capital which minimizes cost and $\bar{\rho}$ the rate of return obtained if output is produced at cost. Note that there are

²³ Williamson [25] argues that large firms have a preference for certain types of expenses, and presents some evidence in support of this contention. These types of expenses would (presumably) be classed as unproductive and therefore excluded from the rate base of a regulated firm.

²⁴ This points to an error in Bailey and Malone [4]. Bailey and Malone assume that the regulatory constraint is binding at the optimal solution and that marginal revenue is strictly positive. From the first-order conditions they obtain $\frac{\phi_K}{\phi_L} = \frac{s}{w}$, and infer from this that at the optimal solution the firm undercapitalizes. This argument fails because at the optimal solution either the regulatory constraint is not binding or marginal revenue is zero; i.e., the argument fails because it involves division by zero. This is essentially the point made by Edelson [10], and it is recognized in Bailey [3] and [5]. However, Bailey [5] argues that the undercapitalization result of Bailey and Malone holds if $\bar{r} \geq s$. This is not correct. It can be seen by inspection of the constraints (27) and (3) that the revenue maximization problem possesses no nontrivial solution if $\bar{r} > s$. The objection raised above applies if $\bar{r} = s$. Bailey [5] and Atkinson-Waverman [1] do not recognize that the revenue maximization problem is not well specified. With one exception noted below in the text, there is an indefinitely large number of *feasible* combinations of capital and labor for each output, and in particular for the output which maximizes revenue. The revenue-maximization problem does not embody any criterion for distinguishing among these feasible combinations; i.e., the optimal values of capital and labor for the problem considered by Bailey and Atkinson and Waverman are indeterminate in *all* but one special case.

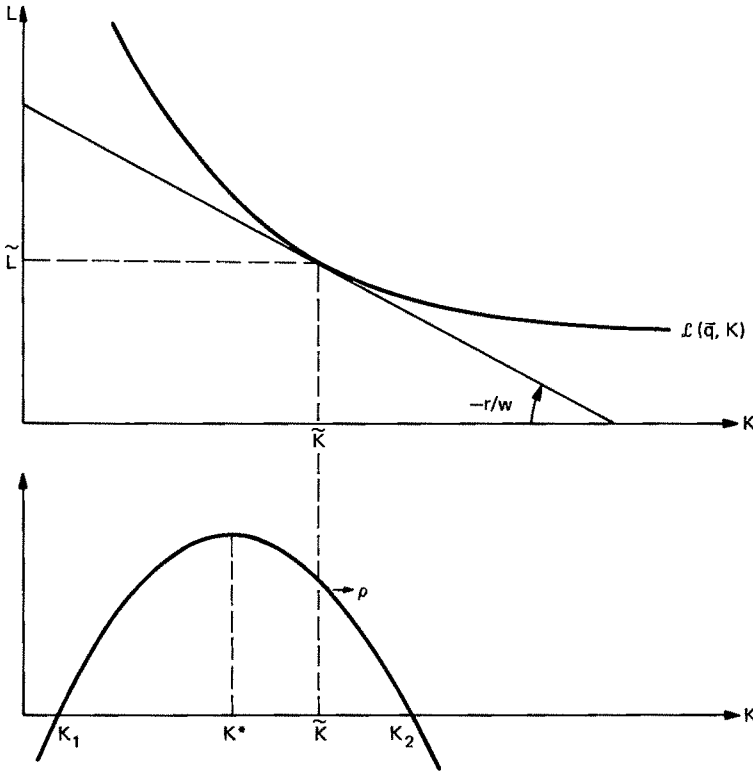
²⁵ In [4].

²⁶ Kafoglis and Bushnell [14] simply point out that the firm may incur "excess production" costs.

²⁷ Bailey [3] presents a proof of a very similar proposition.

FIGURE 4

VARIATIONS IN THE RATE OF RETURN
WITH CAPITAL FOR GIVEN OUTPUT $\bar{q}$



values $K_1 < \tilde{K}$ and $K_2 > \tilde{K}$ for which $\rho = 0$. Since ρ is assumed to be a continuous and strictly concave function of capital for given output and $\bar{\rho} > 0$, there exists a unique value K^* , $K_1 < K^* < K_2$ which maximizes the rate of return at output $\bar{q}$. The question is whether K^* is less than $\tilde{K}$. The first-order condition for $K^* > 0$ to maximize ρ at output $\bar{q}$ is:

$$\frac{\partial \rho}{\partial K^*} = -\frac{1}{K^*} \left\{ \frac{R(\bar{q}, z) - wL(\bar{q}, K^*)}{K^*} + wL_2(\bar{q}, K^*) \right\} = 0, \quad (29a)$$

$$K^* \frac{\partial \rho}{\partial K^*} = -[\rho^* + wL_2(\bar{q}, K^*)] = 0, \quad (29b)$$

where ρ^* is the maximum rate of return. Note that $L_2(\bar{q}, \tilde{K}) = -r/w$, so:

$$\tilde{K} \frac{\partial \rho}{\partial \tilde{K}} = r - \bar{\rho}. \quad (30)$$

It is assumed that the rate of return obtained if cost is minimized decreases monotonically with output and that rate of return r is obtained if q_c is produced at cost. Therefore, $\bar{\rho} > r$ for each output $\bar{q}$, $0 < \bar{q} < q_c$ and hence from (30), $\frac{\partial \rho}{\partial \tilde{K}} < 0$ for $0 < \bar{q} < q_c$. This in turn implies that $K^* < \tilde{K}$, which was to be shown.

FIGURE 5
OPTIMAL OUTPUT OF THE REVENUE MAXIMIZATION PROBLEM

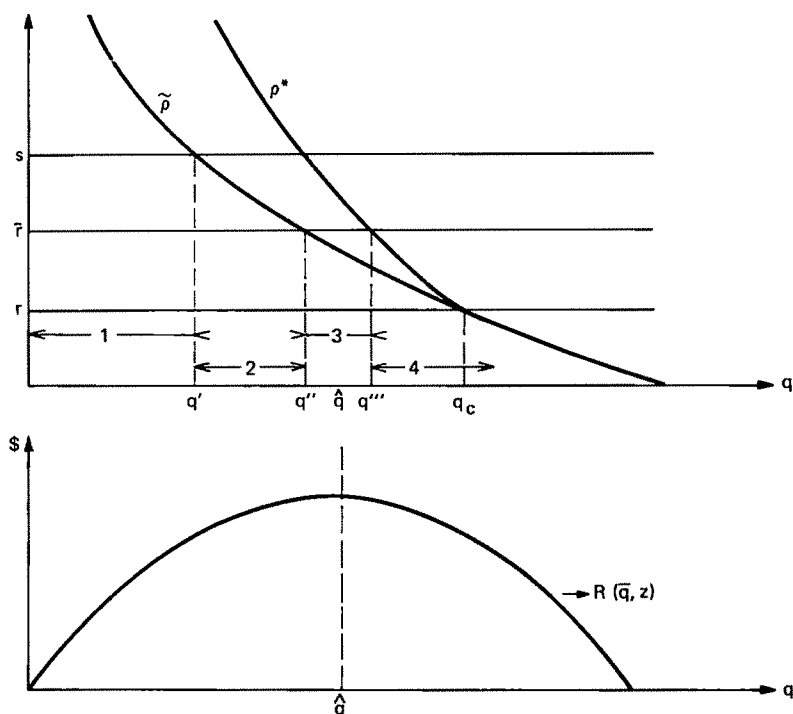


Figure 5 illustrates the significance of this proposition. The curves labeled p^* and $\tilde{p}$ give respectively, the maximum rate of return at each output and the rate of return obtained if output is produced at cost. The largest output which can be produced at cost while satisfying both the regulatory constraint and the profit constraint is q'' and the largest feasible output is q''' . The revenue function assumed in Figure 5 is such that $\hat{q}$ falls between q'' and q''' . If it were forced to minimize cost, the best that the revenue-maximizing firm could do in this case would be to produce q'' . However, by undercapitalizing, it is possible to trade decreased profit for an increased rate of return at each output and thereby it is possible to set output closer to (or as shown, at) $\hat{q}$.

The case shown in Figure 5 is, of course, not the only possibility. The character of the optimal solution clearly depends on the magnitudes of $\hat{q}$, q' , q'' , and q''' . The four possible cases are as follows :

Case 1: $0 < \hat{q} < q'$. It is optimal to produce $\hat{q}$ at minimum expense, which in this case requires overcapitalization.

Case 2: $q' \leq \hat{q} \leq q''$. It is optimal to produce $\hat{q}$ at cost.

Case 3: $q'' < \hat{q} < q'''$. Output $\hat{q}$ is produced with less capital than is required by cost minimization but it is not necessary to maximize the rate of return.

Case 4: $\hat{q} \geq q'''$. The optimal output is q''' , which is feasible if the values of capital and labor are chosen so as to maximize the rate of return.

The comparative statics analysis is facilitated by defining the derived demands that are appropriate for each of these cases. The expense minimizing derived demands for outputs less than q' were

obtained in the preceeding section and the cost minimizing derived demands require no discussion. The remaining task is that of obtaining the derived demands for cases 3 and 4.

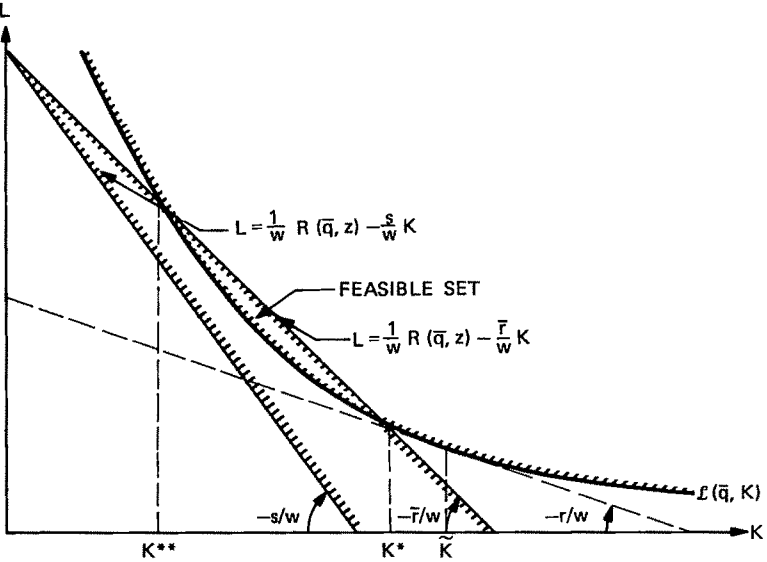
The value of capital which maximizes the rate of return can be obtained by solving (29a) and the corresponding quantity of labor obtained from the labor requirements function. These relationships, which will be referred to as the maximal return derived demands, are denoted $K = H^1(q, w, z)$ and $L = H^2(q, w, z)$.

The maximal return derived demands are defined for all positive outputs, but they are relevant only if $\hat{q} \geq q'''$. If $\hat{q}$ falls between q'' and q''' , then $\hat{q}$ cannot be produced at cost, but it is not necessary to maximize the rate of return. The relevant derived demands in this case are those which yield exactly $\bar{r}$ at minimum expense.

It is clear that undercapitalization is required in case 3, but there is a question as to what input combination minimizes expense. As in the expense minimization problem considered earlier, this can be determined by inspection of the feasible set.

Figure 6 describes the feasible set for an output $\bar{q}$ in the interval

FIGURE 6
THE FEASIBLE SET FOR OUTPUT $\bar{q}$, $q'' < \bar{q} < q'''$



(q'', q''') . The feasible combinations of capital and labor for each output in this interval satisfy:

$$L \leq \frac{1}{w} R(\bar{q}, z) - \frac{\bar{r}}{w} K, \tag{31}$$

$$L \geq \frac{1}{w} R(\bar{q}, z) - \frac{s}{w} K, \tag{8}$$

$$L \geq \mathcal{L}(\bar{q}, K). \tag{6}$$

The largest and smallest feasible values of capital are denoted K^* and K^{**} , respectively, and the input combination which minimizes the cost of producing $\bar{q}$ is denoted $(\bar{K}, \bar{L})$. It is assumed for con-

venience that the allowed rate of return exceeds the maximum rate of return at $\bar{q}$ (i.e., that all points on the isoquant satisfy the rate of return constraint). The profit constraint is, by assumption, not satisfied at $(\bar{K}, \bar{L})$.

It is furthermore assumed that there is a feasible solution, which is to say that the graph of $L = (1/w)R(\bar{q}, Z) - (\bar{r}/w)K$ cannot everywhere be below the isoquant. This graph of L has slope $-\bar{r}/w$, where, since the satisfactory rate of return $\bar{r}$ is stipulated to be less than or equal to r , $|\bar{r}/w| \leq |r/w|$. The graph of L therefore intersects (or is tangent) with the isoquant at a value of capital not less than, and normally greater than $\bar{K}$. Where $\bar{r}$ is strictly less than r , as it is in the graph shown in Figure 6, $K^* < \bar{K}$; i.e., $\bar{q}$ must be produced by using less capital than is consistent with cost minimization.

The profit constraint is satisfied exactly at both K^* and K^{**} . However, given that isoquants are convex to the origin, the expense of producing $\bar{q}$ obviously is minimized subject to the constraint at K^* .

The demands for capital and labor for case 3 can be obtained by the same method used in Section 2. For each output $\bar{q} < q'''$ there are two values of capital for which:

$$R(\bar{q}, z) - [\bar{r}K + w\mathcal{E}(\bar{q}, K)] = 0, \quad \bar{q} < q'''. \quad (32)$$

It is the larger of these two values (denoted K^*) which is relevant. Define $\sigma \equiv R(\bar{q}, z) - [\bar{r}K + w\mathcal{E}(\bar{q}, K)]$. Then

$$\frac{\partial \sigma}{\partial K^*} = \bar{r} + w\mathcal{E}_2(\bar{q}, K^*) = -\frac{1}{K^*} \frac{\partial \rho}{\partial K^*}. \quad (33)$$

Since K^* must exceed the value of capital which maximizes the rate of return at $\bar{q}$,²⁸ $\frac{\partial \rho}{\partial K^*} < 0$, so $\frac{\partial \sigma}{\partial K^*} > 0$. Therefore, (32) can be solved for the value of capital which yields $\bar{r}$ at minimum expense. Substitution of this relationship into the labor requirements function gives the corresponding derived demand for labor. The derived demands so obtained are denoted $K = G^1(q, \bar{r}, w, z)$ and $L = G^2(q, \bar{r}, w, z)$. These derived demands are defined for all positive outputs less than q''' , but they are relevant only for output in the interval (q'', q''') .

□ **Comparative statics.** The regulatory constraint is binding at the optimal solution if case 1 occurs. It is not binding in cases 2, 3, and 4. The issue, then, is whether there are comparative statics properties that distinguish each of these cases from case 1.

The comparative statics properties of derived demand for case 1 (overcapitalization) and case 2 (cost minimization) have already been described. This leaves for consideration the comparative statics properties of derived demand in cases 3 and 4 and the total derivatives of the optimal solution for each of the revenue maximization cases.

²⁸ Since the isoquants are assumed to be convex, for each output $\bar{q} < q'''$, there are values K^* and $K^{**} < K^*$ for which the profit constraint is exactly satisfied. Let $\hat{K}$ be the value of capital which maximizes the rate of return at $\bar{q}$ and suppose that $K^* < \hat{K}$. The rate of return for given output is assumed to be a strictly concave function of capital, so the rate of return decreases monotonically for decreases in capital from $\hat{K}$. Hence, if $K^* < \hat{K}$, there is no value $K^{**} < K^*$ for which the constraint is satisfied, which is a contradiction.

TABLE 3
 PARTIAL DERIVATIVES OF DERIVED DEMAND REVENUE MAXIMIZATION CASES 3 AND 4
 AND TOTAL DERIVATIVES OF THE OPTIMAL SOLUTION, REVENUE MAXIMIZATION CASES 1, 2, 3, AND 4

(A) DERIVED DEMAND	
CASE 3 SATISFACTORY RETURN $q \leq q''$	CASE 4 MAXIMAL RETURN $q > 0$
$\frac{\partial G^1}{\partial q} = \frac{M - wL_1^*}{\bar{r} + wL_2^*} \quad (34a)$	$\frac{\partial H^1}{\partial q} = \frac{wL_1^* - M - K L_{21}^*}{wL_{22}^* K} \quad (35a)$
$\frac{\partial G^1}{\partial \bar{r}} = -\frac{K}{\bar{r} + wL_2^*} \quad (34b)$	$\frac{\partial H^1}{\partial \bar{r}} = \frac{\partial H^1}{\partial r} = \frac{\partial H^1}{\partial s} = 0 \quad (35b)$
$\frac{\partial G^1}{\partial r} = \frac{\partial G^1}{\partial s} = 0 \quad (34c)$	$\frac{\partial H^1}{\partial w} = \frac{L_2^*}{wL_{22}^*} \quad (35c)$
$\frac{\partial G^1}{\partial w} = -\frac{L}{\bar{r} + wL_2^*} \quad (34d)$	$\frac{\partial H^1}{\partial Z_1} = -\frac{\partial R / \partial z_1}{wL_{22}^* K} \quad (35d)$
$\frac{\partial G^1}{\partial z_1} = \frac{\partial R / \partial z_1}{\bar{r} + wL_2^*} \quad (34e)$	$\frac{\partial H^2}{\partial q} = L_1^* + L_2^* \left[\frac{wL_1^* - M - K L_{21}^*}{wL_{22}^* K} \right] \quad (37a)$
$\frac{\partial G^2}{\partial q} = L_1^* + L_2^* \frac{M - wL_1^*}{\bar{r} + wL_2^*} \quad (35a)$	$\frac{\partial H^2}{\partial \bar{r}} = \frac{\partial H^2}{\partial r} = \frac{\partial H^2}{\partial s} = 0 \quad (37b)$
$\frac{\partial G^2}{\partial \bar{r}} = -L_2^* \frac{K}{\bar{r} + wL_2^*} \quad (35b)$	$\frac{\partial H^2}{\partial w} = -\frac{(L_2^*)^2}{wL_{22}^*} \quad (37c)$
$\frac{\partial G^2}{\partial r} = \frac{\partial G^2}{\partial O} = 0 \quad (35c)$	$\frac{\partial H^2}{\partial z_1} = -L_2^* \frac{\partial R / \partial z_1}{wL_{22}^* K} \quad (37d)$
$\frac{\partial G^2}{\partial w} = -L_2^* \frac{L}{\bar{r} + wL_2^*} \quad (35d)$	
$\frac{\partial G^2}{\partial z_1} = L_2^* \frac{\partial R / \partial z_1}{\bar{r} + wL_2^*} \quad (35e)$	
(B) TOTAL DERIVATIVES OF THE OPTIMAL OUTPUT	
CASES 1, 2, AND 3 $0 < q < q'''$	CASE 4 $q''' < q$
$\frac{dq^*}{d\alpha} = 0 \quad \alpha \in (s, \bar{r}, r, w) \quad (38a)$	$\frac{dq^*}{d\alpha} = 0 \quad \alpha \in (s, r) \quad (39a)$
$\frac{dq^*}{dz_1} = -\frac{\partial M^* / \partial z_1}{M'^*} \quad (38b)$	$\frac{dq^*}{d\bar{r}} = \frac{K}{M^* - wL_1^*} \quad (39b)$
	$\frac{dq^*}{dw} = \frac{L}{M^* - wL_1^*} \quad (39c)$
	$\frac{dq^*}{dz_1} = -\frac{\partial R / \partial z_1}{M^* - wL_1^*} \quad (39d)$

Table 3(A) presents formulas for the partial derivatives of derived demand for the two undercapitalization cases. The partial derivatives of $G^1(\cdot)$ are obtained by differentiating (32), and the partial derivatives of $H^1(\cdot)$ are found by differentiating (29a). The corresponding partial derivatives of the demand for labor are in both cases obtained by using the labor requirements function. Table 3(B) presents the total derivatives of the optimal output for each of the revenue maximization cases. It is assumed for simplicity that $\hat{q} \neq q'''$. In each of cases 1, 2, and 3 the optimal output q^* satisfies $M(q^*, z) = 0$. Therefore, q^* is independent of s , $\bar{r}$, r , and w . Differentiation of $M(q^*, z) = 0$ with respect to z_i yields (38b). The regulatory constraint holds with strict equality at the optimal solution in case 4. This fact, and the first-order condition (29b) can be used to obtain the total derivatives of the optimal output. For example differentiation of the profit constraint stated as a strict equality yields on rearrangement:

$$\left[M^* - (\bar{r} + w\mathcal{L}_2^*) \frac{\partial H^{1*}}{\partial q^*} - w\mathcal{L}_1^* \right] \frac{dq^*}{d\bar{r}} - \left[(\bar{r} + w\mathcal{L}_2^*) \frac{\partial H^{1*}}{\partial \bar{r}} + K^* \right] = 0. \quad (40)$$

At the optimal solution in case 4, $\bar{r} = \rho^*$ and, using (29a), $\rho^* + w\mathcal{L}_2(q^*, K^*) = 0$, so (40) implies:

$$\frac{dq^*}{d\bar{r}} = \frac{K}{M^* - w\mathcal{L}_1^*}. \quad (39b)$$

Equations (39c) and (39d) are similarly obtained.

The sign restrictions established are summarized in Table 4. The restrictions on the partial derivatives of $G^1(\cdot)$ and $G^2(\cdot)$ with respect to $\bar{r}$, w , and z_i follow immediately from the formulas presented in Table 3(A), given the signs of $\mathcal{L}_1$, $\mathcal{L}_2$, and $\bar{r} + w\mathcal{L}_2^* > 0$. The corresponding restrictions on the partial derivatives of $H^1(\cdot)$ and $H^2(\cdot)$ also follow immediately from the formulas, given the assumed signs of $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_{22}$. It is assumed that $M' < 0$ for $q \geq 0$, which gives the restriction on $\frac{dq^*}{dz_i}$ shown for cases 1, 2, and 3. The sign restrictions

on the total derivatives of capital and labor for these three cases are obtained immediately using (21a) and (21b). It is necessary to establish the sign of $M^* - w\mathcal{L}_1^*$ to restrict the signs of the total derivatives of the optimal output in case 4. There is undercapitalization at the optimal solution in case 4. Hence, by the same argument used in the overcapitalization case, if $\left(\frac{\partial \phi_L}{\partial K} \right)_q > 0$, $w\mathcal{L}_1(q^*, K^*) > C'$. The monopoly output is assumed to be unique, and the optimal output in case 4 exceeds q_m . It follows that $M^* - w\mathcal{L}_1^* < 0$, which gives the sign restrictions shown. Inspection of (36a) indicates that $\frac{\partial H^{1*}}{\partial q} > 0$

which, together with preceeding results, implies $\frac{dH^{1*}}{d\bar{r}} < 0$. Additional restrictions might be found, but as they do not prove to be crucial this possibility was not explored.

TABLE 4
COMPARATIVE STATICS PROPERTIES OF THE REVENUE MAXIMIZATION MODEL

(A) COMPARATIVE STATICS OF DERIVED DEMAND												
CASES	$\partial K/\partial ()$						$\partial L/\partial ()$					
	q	s	$\bar{r}$	r	w	z_1	q	s	$\bar{r}$	r	w	z_1
OVER CAPITALIZATION $q < q'$	$\pm$	—	0	0	—	$\text{sign } \frac{\partial D}{\partial z_1}$	?	+	0	0	+	$-\text{sign } \frac{\partial D}{\partial z_1}$
COST MINIMIZATION $q \geq 0$	+	0	0	—	+	0	+	0	0	+	—	0
RATE OF RETURN $\bar{r}$ $q \leq q''$	$\pm$	0	—	0	—	$\text{sign } \frac{\partial D}{\partial z_1}$	$\pm$	0	+	0	+	$-\text{sign } \frac{\partial D}{\partial z_1}$
MAXIMAL RATE OF RETURN $q \geq 0$	$\pm$	0	0	0	—	$\text{sign } \frac{\partial D}{\partial z_1}$	$\pm$	0	0	0	—	$\text{sign } \frac{\partial D}{\partial z_1}$

(B) TOTAL VARIATIONS IN THE OPTIMAL SOLUTION															
CASES	$dq^*/d()$					$dK^*/d()$					$dL^*/d()$				
	s	$\bar{r}$	r	w	z_1	s	$\bar{r}$	r	w	z_1	s	$\bar{r}$	r	w	z_1
CASE 1 $q < q'$	0	0	0	0	$\text{sign } \frac{\partial M}{\partial z_1}$	—	0	0	—	?	+	0	0	+	?
CASE 2 $q' \leq \hat{q} \leq q''$	0	0	0	0	$\text{sign } \frac{\partial M}{\partial z_1}$	0	0	—	+	?	0	0	+	—	?
CASE 3 $q'' < \hat{q} < q'''$	0	0	0	0	$\text{sign } \frac{\partial M}{\partial z_1}$	0	—	0	—	$\text{sign } \frac{\partial D}{\partial z_1}$	0	+	0	+	$-\text{sign } \frac{\partial D}{\partial z_1}$
CASE 4 $\hat{q} \geq q'''$	0	—	0	—	$\text{sign } \frac{\partial D}{\partial z_1}$	0	—	0	?	?	0	?	0	?	?

There is little to say about the comparative statics properties of the maximal return derived demands, as these functions are independent of all the parameters of the problem except w and z . The fact that the maximal return derived demands are independent of s , $\bar{r}$, and r does, of course, distinguish them from each of the other possibilities.²⁹

The expense minimizing derived demands for case 3 ($q'' < q < q'''$) have the same peculiar properties found earlier in the overcapitalization case ($q < q'$). In particular, the derived demand for capital *decreases* and the derived demand for labor *increases* with *increases* in the wage rate. This similarity is somewhat surprising, but easily understood. An increase in the wage rate decreases profit [$R - wL$] and the rate of return at the initial input combination. In both the overcapitalization and the undercapitalization cases, the relevant portion of the boundary of the feasible set shifts to the left. However, in the undercapitalization case (Figure 6) the firm is required to become *less* efficient, while in the overcapitalization case it becomes

²⁹ Note that the possibility that the firm maximizes revenue rather than profit can easily be encompassed within an econometric model of the regulated firm. For the example given previously, this would be done simply by including r as one of the independent variables in each of the equations. However, it is not at all clear how $\bar{r}$ should be measured.

more efficient. The appropriate response in both cases is to use less capital and more labor.

The comparative statics properties do nonetheless distinguish overcapitalization from undercapitalization. The regulatory constraint is binding at the optimal solution if and only if the output q which maximizes revenue is less than q' . In this case (overcapitalization), both the partial and the total derivatives of capital and labor are independent of $\bar{r}$ and r but not s . If $q'' < \hat{q} < q'''$ (undercapitalization, ineffective regulation) these derivatives are independent of s and $\bar{r}$ but not, of course, independent of r .

4. Conclusion: profit maximization vs. revenue maximization

■ The preceding sections, by dealing separately with the profit-maximization and the revenue-maximization models, invite two questions. First, do comparative statics properties conclusively distinguish the two models? Second, is it possible to distinguish effective from ineffective regulation without specifying which is the more descriptive model? In fact, there is little need to distinguish these questions, as information which answers one generally answers the other as well.

The comparative statics of derived demand divided the cases which have been considered into three groups: (1) overcapitalization ($0 < q < q'$); (2) cost minimization ($q' \leq q \leq q''$); and (3) undercapitalization ($q > q''$). The firm will never undercapitalize if it maximizes profits, so the question of whether the firm maximizes revenue or profit does not arise in the last of these cases.

Comparison of Tables 2 and 4 indicates that it is possible to distinguish profit maximization from revenue maximization whether or not the regulatory constraint is binding. If the firm maximizes profit, then the optimal output decreases with the allowed rate of return if regulation is effective or decreases with the competitive cost of capital if regulation is ineffective. If, instead, the firm maximizes revenue, the optimal output is independent of both the allowed rate of return and the competitive cost of capital.

The question of whether the firm maximizes profit or revenue would be particularly interesting if observation were consistent with the hypothesis that cost is minimized. In this case, it is relevant to ask why the constraint is not binding—because (1) the firm maximizes profit and the allowed rate of return exceeds the rate of return at the monopoly solution, or (2) because the firm maximizes revenue and the allowed rate of return is such that $\hat{q}$ can be produced at cost? If the first of these possibilities is in fact correct, the regulatory agency presumably should reduce the allowed rate of return. However, no action by the regulatory agency is warranted if the regulatory constraint is ineffective because revenue is maximized.

Appendix

■ 1. It is pointed out in the text that an output is not produced at minimum expense if inputs are wasted. Furthermore, for outputs in the interval $[0, q']$ the regulatory constraint is satisfied as a strict equality at the input combination that minimizes expense. Therefore, the problem can be stated as:

$$\min_{\{K\}} [E] = rK + w\mathcal{L}(q, K) - \theta[sK + w\mathcal{L}(q, K) - R(q, z)], \quad (41)$$

Properties of the marginal expense function

where θ is an undetermined Lagrangian multiplier. The first-order conditions for $K^* > 0$ and θ^* to minimize E are:

$$r + w\mathcal{L}_2 - \theta^*(s + w\mathcal{L}_2) = 0, \quad (42)$$

$$R(q, z) - [sK + w\mathcal{L}(q, K)] = 0. \quad (43)$$

The constraint is binding ($\theta^* > 0$) for $0 \leq q < q'$ and is not binding ($\theta^* = 0$) for $q \geq q'$. The second-order condition for a regular minimum is:

$$(1 - \theta^*)\mathcal{L}_{22}(q, K^*) > 0. \quad (44)$$

It is assumed that $\mathcal{L}_{22} > 0$ for $q > 0, K > 0$. This assumption and inequalities (43) and (44) then require $0 \leq \theta^* < 1$ for $\nu > 0$ and $0 < q \leq q'$. Note also that as (41) can be restated as $(1 - \theta^*) \times (r + w\mathcal{L}_2) = \nu\theta^*$, this restriction on θ^* implies $r + w\mathcal{L}_2 > 0$ for $\nu > 0$ and $0 < q < q'$. Equation (42) is satisfied at each of points b^* and b^{**} of Figure 1. If $\nu = 0$, (41) is also satisfied at each of these points for $\theta^* = 1$, but inequality (44) is not satisfied, which is appropriate, as the optimal solution is not unique.

2. Marginal expense is:

$$\mathcal{E}' = (r + w\mathcal{L}_2) \frac{\partial E^1}{\partial q} + w\mathcal{L}_1, \quad \nu > 0, \quad 0 < q \leq q'. \quad (45a)$$

The combination $(\tilde{K}, \tilde{L})$ required to produce q' at cost is feasible, and $r + w\mathcal{L}_2(q', \tilde{K}) = 0$. Therefore:

$$\mathcal{E}'(q', s, w, z) = w\mathcal{L}_1(q', \tilde{K}) = \frac{w}{\phi_L(\tilde{K}, \tilde{L})} = C'(q', r, w), \quad (46)$$

where $C'(q, r, w)$ is marginal cost.

3. For any output $\bar{q} < q'$, let $(\tilde{K}, \tilde{L})$ be the combination of capital and labor required to produce $\bar{q}$ at cost and (K^*, L^*) the combination required to produce at $\bar{q}$ at minimum expense subject to the regulatory constraint. Let $\tilde{\phi} = \phi(\tilde{K}, \tilde{L})$ and $\phi^* = \phi(K^*, L^*)$, and note that marginal cost is $C'(\bar{q}, r, w) = w/\tilde{\phi}$. It has been established that if $\nu > 0$ the firm uses more capital and hence less labor than is consistent with cost minimization at each output $q < q'$. Therefore $\phi^* > \tilde{\phi}$ if the marginal productivity of labor increases with increases in capital along an isoquant. This will be the case if and only if:

$$w\mathcal{L}_{12} = - \frac{1}{\phi_L^2} [\phi_{LK} + \mathcal{L}_2\phi_{LL}] < 0. \quad (47)$$

The bracketed term in this expression is the rate of change of the marginal productivity of labor with respect to capital, given output. If, as would typically be assumed, $\phi_{LK} > 0$ and $\phi_{LL} < 0$, then $\left(\frac{\partial\phi_L}{\partial K}\right)_q > 0$. Of course, it still may be that $\left(\frac{\partial\phi_L}{\partial K}\right)_q > 0$, and hence $w\mathcal{L}_{12} < C'$, although either $\phi_{LK} < 0$ or $\phi_{LL} > 0$ (but not both).

4. The form of the relationship between marginal expense and marginal revenue can be easily derived. Differentiation of (10a)

yields on rearrangement:

$$\frac{\partial E^1}{\partial q} = \frac{M - w\mathcal{L}_1^*}{s + w\mathcal{L}_2^*}, \quad v > 0, \quad 0 < q \leq q', \quad (48)$$

where $M = \frac{\partial R}{\partial q}$. Substitution of (44) into (41a), then gives

$$\mathcal{E}' = \left[\frac{r + w\mathcal{L}_1^*}{s + w\mathcal{L}_2^*} \right] (M - w\mathcal{L}_1^*) + w\mathcal{L}_1^*, \quad v > 0, \quad 0 < q \leq q'. \quad (45b)$$

Since, from (41), $\theta^* = \frac{r + w\mathcal{L}_2^*}{s + w\mathcal{L}_2^*}$, this simplifies to:

$$\mathcal{E}' = \theta^* M + (1 - \theta^*) w\mathcal{L}_1^*, \quad v > 0, \quad 0 < q \leq q'. \quad (45c)$$

References

1. ATKINSON, A. AND WAVERMAN, L. "Resource Allocation and the Regulated Firm: Comment." *The Bell Journal of Economics and Management Science*, Vol. 4, No. 1 (Spring 1973), pp. 283-287.
2. AVERCH, H. AND JOHNSON, L. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol. 52, No. 5 (December 1962), pp. 1053-1069.
3. BAILEY, E. "Economic Theory of Regulatory Constraint." Unpublished doctoral dissertation, Princeton University, 1972.
4. ——— AND MALONE, J. "Resource Allocation and the Regulated Firm." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 1 (Spring 1970), pp. 129-142.
5. ———. "Resource Allocation and the Regulated Firm: Comment on the Comments." *The Bell Journal of Economics and Management Science*, Vol. 4, No. 1 (Spring 1973), pp. 288-292.
6. BAUMOL, W. *Business Behavior, Value and Growth*. New York: The Macmillan Company, 1959 (rev. ed., 1967).
7. ——— AND KLEVORICK, A., "Input Choices and Rate of Return Regulation: An Overview of the Discussion." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 162-190.
8. COMANOR, W. "Should Natural Monopoly Be Regulated?" *Stanford Law Review*, Vol. 22 (February 1970), pp. 510-518.
9. DAYAN, D. "Vertical Integration and Monopoly Regulation." Unpublished doctoral dissertation, Princeton University, 1972.
10. EDELSON, N. "Resource Allocation and the Regulated Firm: A Reply to Bailey and Malone." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 374-378.
11. EL-HODIRI, M. AND TAKAYAMA, A. "Behavior of the Firm under Regulatory Constraint: Clarifications." *The American Economic Review*, Vol. 63, No. 1 (March 1973), pp. 235-237.
12. JACKSON, R. "Regulation and Electric Utility Rate Levels." *Land Economics*, Vol. 35, No. 3 (August 1969), pp. 372-376.
13. KAFOGLIS, M. "Output of the Restrained Firm." *The American Economic Review*, Vol. 59, No. 4 (September 1969), pp. 583-589.
14. ——— AND BUSHNELL, R. "The Revenue Maximization Oligopoly Model: Comment." *The American Economic Review*, Vol. 60, No. 3 (June 1970), pp. 427-428.
15. KLEVORICK, A. "The 'Optimal' Fair Rate of Return." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 122-153.
16. MACAVOY, P. "The Formal Work-Product of the Federal Power Commissioners." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 379-395.
17. ———. *Price Formation in Natural Gas Fields*. New Haven: Yale Univ. Press, 1962.

18. MOORE, T. "The Effectiveness of Regulation of Electric Utility Price." *Southern Economic Journal*, Vol. 36, No. 4 (April 1970), pp. 365-375..
19. PRESSMAN, I. AND CAROL, A. "Behavior of the Firm under Regulatory Constraint: Note." *The American Economic Review*, Vol. 61, No. 1 (March 1971), pp. 210-212.
20. ———. "Behavior of the Firm under Regulatory Constraint: Reply," *The American Economic Review*, Vol. 63, No. 1 (March 1973), p. 238.
21. SHESHINSKI, E. "Welfare Aspects of a Regulatory Constraint: Note." *The American Economic Review*, Vol. 61, No. 1 (March 1971), pp. 175-178.
22. STIGLER, G. AND FRIEDLAND, C. "What Can Regulators Regulate? The Case of Electricity." *Journal of Law and Economics*, Vol. 5 (October 1962), pp. 1-16.
23. TAKAYAMA, A. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol. 59, No. 3 (June 1969), pp. 255-260.
24. WESTFIELD, F. "Regulation and Conspiracy." *The American Economic Review*, Vol. 55, No. 3 (June 1965), pp. 424-443.
25. WILLIAMSON, O. *The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm*. Englewood Cliffs, N. J.: Prentice-Hall, 1964.
26. ZAJAC, E. "A Geometric Treatment of Averch-Johnson's Behavior of the Firm Model." *The American Economic Review*, Vol. 60, No. 1 (March 1970), pp. 117-125.
27. ———. "Note on 'Gold Plating' or 'Rate Base Padding.'" *The Bell Journal of Economics and Management Science*, Vol. 3, No. 1 (Spring 1972), pp. 311-315.

Alternative regulatory policies for dealing with the natural gas shortage

Paul W. MacAvoy

Professor of Economics
Sloan School of Management
Massachusetts Institute of Technology

and

Robert S. Pindyck

Assistant Professor of Economics
Sloan School of Management
Massachusetts Institute of Technology

Low wellhead ceiling prices over the past decade have led to the beginning of a shortage in natural gas production. If the demand for gas grows as expected during the 1970s, and if ceiling prices remain low as a result of restrictive regulatory policy, this shortage could grow significantly. This paper examines the effects of this and alternative regulatory policies on gas reserves, production supply, production demand, and prices over the remainder of this decade. An econometric model is developed to explain the gas discovery process, reserve accumulation, production out of reserves, pipeline price markup, and wholesale demand for production on a disaggregated basis. By simulating this model under alternative policy assumptions, we find that the gas shortage can be ameliorated (and after four or five years, eliminated) through phased deregulation of wellhead sales, along the lines proposed in President Nixon's April 1973 Energy Message. But the shortage is not reduced by changes in regulatory rulings on prices, even when the rulings allow price increases that are the maximum likely to be acceptable to the Courts. These results are rather insensitive to alternative forecasts of such exogenous variables as GNP growth, population growth, and changes in the prices of alternate fuels.

1. Introduction

■ A substantial shortage of natural gas has been developing over the last few winter heating seasons. Reports of the Federal Power Commission indicate that in the winters of 1970 to 1972 gas supplies were cut off with increasing frequency, and for longer periods of time, throughout the Northern and Eastern portions of the United States.¹

Paul W. MacAvoy holds the A.B. degree from Bates College (1955) and the M.A. and Ph.D. degrees from Yale University (1956 and 1960, respectively). At the Sloan School of Management, his research centers on the effects of regulation in the energy, transportation, and communications industries. Professor MacAvoy is Editor of *The Bell Journal of Economics and Management Science*.

Robert S. Pindyck received the B.S. and the M.S. in electrical engineering and the Ph.D. in economics from M.I.T. in 1966, 1967, and 1971, respectively. At the Sloan School of Management, he conducts research in the economics of the energy sector, as well as in applications of optimal control theory to macroeconomic and industry policy formation.

¹ In [21].

This has been a matter not only of cutting off supplies in peak periods to industry—as occurred in Cleveland in January 1970 when 30,000 employees of 700 companies were laid off for 10 days as a result of gas interruptions—but of systematic curtailments of deliveries to certain classes of consumers. During 1971 to 1972, seven major interstate pipelines curtailed service throughout the winter season to existing customers. Service entering the Northeastern part of the country from Texas Eastern Transmission Corporation was curtailed by 18 percent, from Transcontinental Gas Pipeline by 9 percent, and from Trunkline Gas Company by 27 percent. Service throughout the region from New Mexico to Southern California was curtailed by El Paso Natural Gas Company by 15 percent. The FPC staff has shown deliveries falling short of the amount of gas demanded for consumption by 3.6 percent in 1971 and by 5.1 percent in 1972, and has predicted that production will fall short of demand by 12.1 percent by 1975.²

These estimates encompass recent and expected future *production* shortages. The amounts do not, however, include anywhere near all of the unsatisfied demands for gas. The “markets” for natural gas are not spot markets for production, but rather center on sales of reserves. The pipeline buyer offers to take new gas deposits out of the ground, and pays for this production, so as to deliver gas to residential, commercial, and industrial consumers throughout the United States. The transaction involves the dedication of reserves by oil and gas discovery companies to pipelines, and this is followed by a transaction in which the pipelines make commitments to deliver production to final consumers over a ten- to twenty-year period. The reserves markets have experienced shortages to a much greater extent than shown by recent production shortages (according to one

This paper reports on the results of Phase II of a project to develop an Econometric Policy Model of Natural Gas, under National Science Foundation Grant #GI-34936. The Phase I version is reported in Sloan School of Management Working Paper #635-72 (December 1972), and the Phase III version, a Domestic United States Oil and Gas Policy Model, will be completed by July 1974. Comments are invited on this interim version.

The results of this project would not have been possible without the support of our research assistants at M.I.T. We would like to extend our sincere appreciation to Robert Brooks, Krishna Challa, Ira Gershkoff, Kevin Lloyd, Marti Subrahmanyam, and Philip Sussman for their energetic and enthusiastic help in developing a computerized data base during the summer of 1972, and in estimating and simulating the model over the past few months. They worked many long hours, and often under considerable time pressure. We received considerable support from the National Bureau of Economic Research Computer Center in the use of the TROLL system for the estimation and simulation of the model, as well as for the maintenance of our data base. We would like to thank Mark Eisner, Fred Ciarmaglia, Richard Hill, and Jonathan Shane for their help in using TROLL. We would also like to express our appreciation to Morris Adelman and Gordon Kaufman of M.I.T., and Robert Fullen and Wade Sewell of the Federal Power Commission for their comments and suggestions. Commissioner Nassikas of the FPC, Mr. Charles DiBona of the White House Staff, and staff members of the Senate Commerce Committee were of substantial assistance in formulating policy alternatives. Finally, more than six dozen readers of the Phase I version of the model provided critical comments necessary for reformulating the model along the lines presented below. We are obligated to them all, and absolve them from responsibility for the results by granting them anonymity.

² In [20], p. 123.

estimate, new reserves have been more than 50 percent short since 1961³). Also, it has been estimated that demands for new production by old and new potential buyers exceeded the total supply of new production available by more than 50 percent.⁴ These unsatisfied demands were never registered by curtailment proceedings because most of the new potential buyers simply were refused service from the pipelines. This shortage resulted in additional demands for other fuels.

Production curtailments have immediate effects on the consumer. They may require him to use stand-by heating facilities or possibly to do without heating entirely. The larger shortage of new production contract commitments has more diverse effects. Some consumers are required to go to other fuels—with the result that their consumption costs are increased or that pollution is greater, so that social costs are increased. Other consumers can work their way around the shortage by changing locations or by offering various premia to be put at the head of the queue. In general, the economic and locational effects of the shortage are likely to be significant.⁵

The political consequences are another matter of import. The contract sales of gas at the wellhead and at the city gate are closely regulated by the Federal Power Commission, so that the FPC, the Congress, and the Office of the President become focal points for complaints that regulatory policies have either caused or failed to ameliorate the shortages. With neither producers nor consumers supporting the regulatory process—neither clearly benefiting from the shortage—the possibility of significant political losses for legislators or regulators is substantial.

Reactions to the potential adverse political effects have been along two lines. The first has been to call for the loosening of regulation of contractual agreements at the wellhead. President Nixon in April 1973 called for deregulation of wellhead prices on new contracts; the Chairman of the FPC at the same time argued that “gas supplies are short . . . and the way to encourage more drilling and discoveries

³ The estimate was obtained from a four-equation supply and demand model of natural gas reserves constructed on the basis of data over the 1950s and fitted to values of the exogenous variables over the 1960s. The difference between the fitted values of reserves—those that would clear the market—and the actual values of reserves made available in fuel producing markets to go to the East Coast and Midwest was more than 50 percent of fitted reserves each year from 1961 to 1968. See Breyer and MacAvoy [5], pp. 968–976.

⁴ *Ibid.*, Table 1, p. 975. This is confirmed in Tables 3, 4, and 5 below, by the year-to-year differences in supply of production falling short of the year-to-year differences in demands for production by more than 50 percent in the 1971 to 1973 period.

⁵ An initial attempt is made by Breyer and MacAvoy [5] to measure the economic effects of the shortage on consumers. It is argued there that the reserve shortage is incurred by residential and commercial consumers buying from regulated interstate pipelines, while industrial consumers buying in unregulated transactions do not experience the shortage to the same extent. Because of the magnitude of the amounts short in the 1962 to 1968 period, final residential and commercial consumers are believed to have been made worse off as a group from a combination of lower regulated prices and large shortages. No attempt is made here to refine these estimates with the more advanced econometric model discussed below. But additional work along these lines—leading to an assessment of optimal levels of shortage in terms of economic effects—will be forthcoming in a later article.

may be to let prices rise.”⁶ Proposals for deregulation are based on the argument that FPC rulings have restricted price increases, despite the fact that cost increases have reduced supplies at the same time that demand has increased. Thus, decontrol would allow higher prices which would clear the market of excess demand and would be an inducement to take on increased exploration and development so as to add to reserve supply.

The second reaction has been to the opposite effect. Proposals have been made to put stronger controls over wellhead contracts. The argument, in part at least, is that producers have been holding back reserves in anticipation of relaxed price controls—and tighter controls would cause them to see the futility of holding back. Draft bills proposed by staff of the Senate Interior and Commerce Committees extend FPC jurisdiction to cover not only interstate sales to pipelines, but all intrastate sales now outside the jurisdiction of the Federal Power Commission. The requirement that the FPC set “just and reasonable rates” on the basis of historical average costs of exploration and development is reaffirmed. Proposals are made to further the development of artificial gas or liquefied natural gas to replace the short supplies of inground natural gas reserves within the domestic United States. In effect, prices are held constant, and price controls are extended to encompass all relevant sources of supply and demand under this policy.

Another rationale for strengthened regulation is that, “given the relatively large unsatisfied demand for gas, deregulation of natural gas prices would lead to massive increases in wellhead prices to abnormally high levels.”⁷ There would be little increase in supply—“there is evidence that gas supplies are relatively inelastic in the short run.”⁸ There is also asserted to be some basis for arguing that present regulation is not the cause for the present shortage. At least, “although cost based regulation was slow in getting started, it is an adequate method of regulation that has been developed in the 1960s. Many of the uncertainties have been worked out. . . . The system of regulation can meet the needs of the 1970s to elicit the necessary supply of natural gas at the lowest reasonable price.”⁹

These proposals will be evaluated in this paper. After a more explicit rendition of alternative policies, the industry into which these policies are to be introduced is discussed in some detail. Then an econometric model for policy analysis is described and, finally, the model is used to evaluate the policy options. This evaluation is in terms of the extent to which the alternative options are likely to ameliorate the shortage.

□ **Three policy alternatives.** Frequent assessments of the size of the gas shortage bring forth even more frequent changes in Congress and the Office of the President in proposals for dealing with the gas shortage. Each of these new policies could be catalogued and evaluated—but the results would be so specific as to be relevant only to that policy, and the duration of the policy might well be rather short.

⁶ See “F.P.C. Head Urges End to Gas Curbs,” *The New York Times*, April 11, 1973, p. 19.

⁷ U. S. Senate Commerce Committee [23], p. 4.

⁸ *Ibid.*

⁹ *Ibid.*, p. 5.

Alternatively, the policies can be characterized along two or three dimensions; this is attempted here, in the expectation that the characterizations will approximate the many specific alternatives to be considered, accepted, or rejected by the Government in the coming three or four years.

The first alternative policy is that of deregulation of wellhead prices of natural gas. The Federal Power Commission sets regional limits on prices for gas dedicated to interstate pipelines; these limits would be loosened or eliminated in a number of specific policy proposals. In general, the FPC price ceilings would be eliminated only after a number of years, by taking controls only off prices in all new contracts signed in each year (the time lag in price increases would be extensive, since most of the flowing gas is under old contracts signed in previous years). Even these prices would not be free to rise to a level that would clear out all the excess demands immediately, since most proposals—including President Nixon's Energy Message of April 1973—decree that there be some national ceiling imposed on new prices in keeping with gradual elimination of the deficit and with antiinflationary price controls. The gradual loosening of price controls is coupled with "short-term" rationing schemes that may involve the extension of regulatory jurisdiction to presently unregulated companies that are bidding up prices. This class of policies can be evaluated in terms of (1) elimination of price controls under new contracts, (2) an overall price ceiling in keeping with a 50-percent increase in *average* wholesale price over five years, and (3) some regulatory jurisdiction over all field sales by the Federal Power Commission.

The second alternative class of proposals centers on more rather than less regulation by the FPC. Prices would be set regionally or nationally on the basis of "cost of service." The cost of service is found by Commission and Court judgment of historical records on average exploration and development costs in the region. All production in the region would come under the jurisdiction of the Commission (except perhaps for the smallest producers, who would be exempt in order to cut down on the number of case reviews). Production under these conditions would admittedly be short of demands; holding prices to historical averages essentially limits reserves or production to amounts equal to or less than historical levels, and these historical levels were not sufficient to meet demands in the past. As a result, attention has to be centered on inducements under regulation for the development of alternative gas supplies, either through manufacture or import in liquefied form. These alternative supplies would be regulated as well. As a result, the second class of policies would (1) set price ceilings close to 1972 to 1973 price levels, with perhaps a one cent per annum increase as historical average costs rise slowly with inflation, (2) place all dedications of reserves and production of gas under the jurisdiction of the FPC, and (3) set the regulated price on manufactured or liquefied gas according to the specific cost of providing these products.

These first two alternatives are basically contradictory, and it does not seem likely that amendments could be made to either that would capture the loyalty of those supporting the other. If neither can be made politically effective, then the relevant alternative—the third—is the status quo, which consists of regulation according to

rulings in recent cases at the Federal Power Commission. In recent years the Commission has followed policies of allowing field price increases in keeping with "changes in average historical costs;" but wide variation in the estimate of such costs has occurred—and the Commission has probably chosen the estimate to be as large as permissible by Court review. Area ceilings reached by negotiated settlements between producers and consumers have been proposed to the Commission, and the Commission has found them to be "reasonable ceiling prices" not outside the range of possible average costs.¹⁰ Comparisons with costs could be made in the near future that would justify higher future prices, if the Commission so willed—comparisons of cost on the most recent contracts for sale of intrastate (unregulated) gas, rather than of interstate (regulated) gas sold over the last few years. The Commission, in the 1970 to 1973 period, allowed prices on new contracts to increase each year by two-and-one-half cents per thousand cubic feet, after having held prices to one-half cent increases each year over the 1960s; the regulator can continue to find estimates of historical average costs that would allow further relatively large price increases. But there are limits to the extent of price increases in keeping with *some* estimate of costs—Court appeals by buyers could lead to rejection of FPC decisions unless there were some justification on cost grounds given in the decisions. Thus, the third class of policies, which can be characterized as maintaining the status quo, includes (1) prices that increase from two to four cents per Mcf in each of the next five years, (2) limited regulatory jurisdiction over intrastate sales, and (3) manufactured and liquefied gas with regulated price ceilings in keeping with their respective costs.

■ In order to analyze the effects of these alternative policies, a model of natural gas markets with explicit policy controls has been constructed. The model provides a vehicle for performing simulations using different policy options, so as to indicate the effects of the options on the size of shortages. Thus, its formulation stresses prices, reserve quantities, production quantities, and associated demands for production.

The model, which will be described in more detail in the next section, treats simultaneously the field market for reserves (gas producers dedicating new reserves to pipeline companies at the wellhead price) and the wholesale market for production (pipeline companies selling gas to retail utilities and industrial consumers). The linking of these two markets by interstate pipeline companies is an important characteristic of the natural gas industry. Delivery in the wholesale market is a determinant of pipelines' demands for gas reserves in the field market, and wholesale delivered prices are determined in part by the prices on new reserve contracts in the field market.

In surveying the interaction of these two markets, we shall consider their behavior in the rather complex case when there are excess demands for production. Pricing policies, under both regulation and deregulation, are briefly reviewed for their impact on the size of the

2. Characteristics of an economic model of natural gas reserves and production

¹⁰ U. S. Federal Power Commission [22]. See also FPC [19].

shortage. Pricing will be quantitatively analyzed in the econometric model in Section 3 and evaluated in the simulations in Section 4.

□ **Field markets.** These markets are the locus of transactions between oil and gas producers having volumes of newly discovered reserves and pipeline buyers seeking to obtain by contract the right to take production from these reserves. The amounts of reserves committed by the oil and gas companies are based primarily on the amounts of inground deposits of oil and gas. Additions to reserves come about from additions in (1) gas associated with newly discovered or developed oil reserves—"associated" gas reserves—and (2) gas volumes found in reservoirs not containing oil—or "non-associated" reserves; both result from new discoveries and from extensions of previous discoveries or revisions of earlier estimates of previous discoveries. The amounts actually in place in the producing reservoirs limit the amounts of both associated and nonassociated new gas reserves that can be "supplied" or dedicated to the pipelines.

Needless to say, there is a finite amount of gas in the ground that can be discovered, and thus we should observe a depletion effect on volumes dedicated. It is not our objective to predict how much gas there actually is remaining to be discovered, or when this finite resource will be depleted. But an economic model of the gas industry should at least take this depletion effect into account; we shall do this by using indicators of depletion or of decreasing returns that occurred over the past decade as variables explaining "supply."

There are important economic determinants of the amount of reserves available for commitment as well, and these include prices of new contracts signed by producers, expected future prices under possibly forthcoming new contracts in subsequent years, and changes in average and marginal costs of exploration and development. These factors are widely believed to have substantial effects upon reserve availability, although with considerable time lags. Commitments today to higher prices for new contract volumes might lead to immediate increased planning activities for further exploratory or developmental work; this might lead in a year or two to additional drilling activity and, in a subsequent year or two, to the offering of additional reserves for sale to pipeline buyers.

Production of gas into the pipelines, termed short-run or annual supply, is limited by the amount of reserves available. Production cannot take place at rates greater than some fixed percent of reserves per annum because sandstone contained in the reservoirs is impermeable or because faster rates by "channeling" and sealing off parts of the reservoir may reduce the economic value of the remaining reserves. But up to that limit, more production in this period can take place at higher short-run costs. This is justified if prices are higher. Thus, for both technical and economic reasons, the supply of production out of reserves will be less the lower is the volume of reserves and the lower is the price in the contract commitment.

The demands for new reserves in the field market are manifest in the willingness to buy on the part of pipeline companies and local (direct) consumers. They seek to obtain long-term contracts for the extraction of the reserves. Demand determinants in the field markets include the wellhead price that pipelines and others are willing to pay for additions to their reserve holdings, the amount of reserves

available at a specific location, the location of these reserves, and the final demands for new production by the buyers repurchasing the gas from the pipelines. These demands would be "registered" in the market only so long as there was no regulatory price control which set regional field prices below those at which total demands were equal to the available supplies of new reserves. After 1961, when ceiling prices were put into effect, the demand variable became the exogenous FPC-determined price, since it was lower than the price that would have equilibrated markets.

These characteristics of field markets are found in many futures markets in which raw materials are dedicated for production and refining. The important characteristics of these particular markets are that (1) more will be made available for sale if the buyers offer higher prices (in rough approximation to the competitive "supply" mechanism), (2) the lag adjustment process bringing forth additional supplies of reserves is likely to be long and possibly quite complex, (3) production out of reserves is determined by a combination of technical and economic circumstances, but is likely to be greater the larger the volume of reserves available and the higher the contract prices pipelines are paying for the gas they are taking, and (4) demands depend on prices, but are also derived from final residential, industrial, and commercial consumption.

□ **Wholesale markets.** Pipelines provide gas deliveries, usually under long-term contract, to industrial consumers taking gas right off the line and to retail public utility companies for resale to industrial, residential, and commercial consumers. The amounts of gas demanded by direct (mainline) industrial consumers and retail gas utilities are believed to depend upon the prices for wholesale gas contracts, the prices for alternative fuels consumed by final buyers, and economy-wide variables that determine the overall size of energy markets. If the consumers are industrial companies seeking gas as boiler fuel or process material, the "market size" variables relate to the demands for these companies' outputs and to their investment in capacity to burn fuel or to utilize energy. If the final consumers are households and commercial building owners, then the "market size" variables relate to total population and income.

Wholesale markets also operate with lags. Changes in wholesale prices quoted by pipeline sellers of gas production feed through as changes in final consumer prices and then feed back as changes in final consumer demand quantities; the feedback may take some time because of consumers' commitment to gas burning equipment and the necessity for that equipment to wear out before demands are reduced. As a result, the effects of price changes may be felt only in reduced demands for more gas in subsequent years.

The amount of production provided to industrial and public utility buyers is not determined by a fixed "supply schedule" of quantities at various prices. The pipelines offer service in the form of delivered gas supplies at a markup over the field purchase price for the reserves backing up that production. The price markups are determined by the cost of transmission and add-on profits limited by FPC regulation (following orthodox public utility procedures of finding cost of capital by taking a "fair return" on "fair value" of original investment outlay for pipeline equipment). They vary widely

in level among wholesale markets—there are one to four sources of transmission capacity in most wholesale buying regions—but in most regions the process seems to be to set the markup and then to provide production out to the limit of either reserves or pipeline capacity.

Markup pricing has been formulated to build in significant lags from changes in field contracts. The Commission has followed the policy of allowing the markup for historical average transmission costs plus profit for the pipelines over historical average field prices for gas at the wellhead. This “rolled-in” price at wholesale is thus changed by an increased field price only to the extent that new contract prices change the historical average of all field prices.¹¹ The full impact of a field price change on wholesale prices would occur only after that change had been in effect for roughly ten years (assuming 10 percent of contracts in each year are new contracts).

There is considerable simultaneity in the behavior of production, reserves, and prices in field and wholesale markets. Field prices determine the availability of new reserves and the production conditions under new contracts at the same time. Changes in field prices are reflected, albeit slowly, in changing wholesale prices and demands for quantities of production at wholesale. This two-level industry can thus be modeled by simultaneous equations estimates of production and prices as they depend upon reserves and conditions in the final markets for energy.

□ **The behavior of gas markets when there are shortages.** The behavior of this “mixed” set of markets—some of which exhibit competitive characteristics of supply while others follow noncompetitive patterns—may be rather complex when there are significant excess demands for production. Structural equations, defined to delineate behavioral patterns, are discussed in some detail later, but for now let us examine how shortages resulting from regulatory policy move through the different layers of transactions for gas reserves and production.

In the field market for natural gas, proved reserves are incremented through the discovery process and depleted through production. The amount of additions to reserves is positively related to some extent to the field price of gas contracts being signed at the locations close to where that volume of new reserves is available. The demands for new reserves by pipelines could be specified in terms of

¹¹ The effects of rolling in a “one-shot” price increase on new contracts in 1974 can be spelled out in detail. In 1973, the wholesale price is just the wellhead ceiling price P_c plus the pipelines’ markup M_t . In 1974, after the ceiling price on new contracts has been raised to P'_c , the wholesale price to buyers of new contracts would *not* be $P'_c + M_t$, but would instead be given by:

$$PW_{1974} = P_c + M_t + \frac{\delta Q_{1974}}{Q_{1974}} (P'_c - P_c),$$

where $\delta Q/Q$ is the proportion of “new” production to total production. In 1975, the wholesale price would be given by:

$$PW_{1975} = P_c + M_t + \frac{\delta Q_{1975}}{Q_{1975} - \delta Q_{1974}} (P'_c - P_c),$$

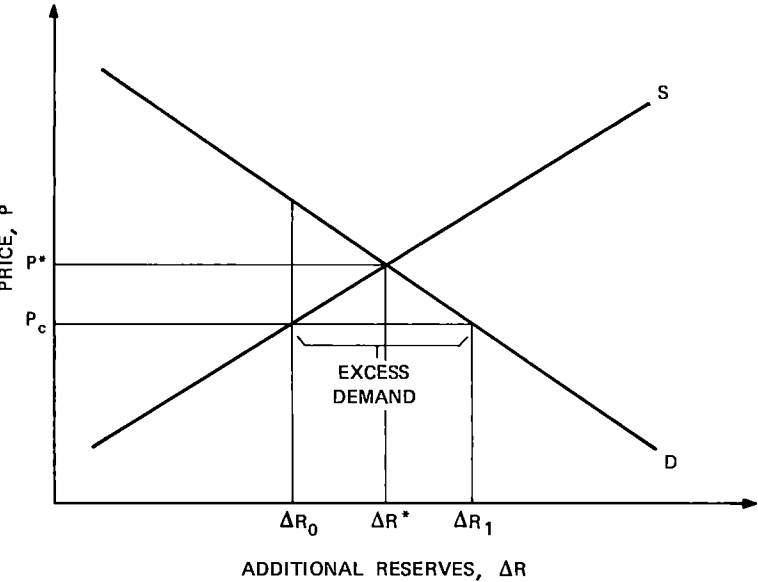
and in 1976, it would be given by:

$$PW_{1976} = P_c + M_t + \frac{\delta Q_{1976}}{Q_{1976} - \delta Q_{1975} - \delta Q_{1974}} (P'_c - P_c).$$

The wholesale price would continue to rise until $\delta Q/(Q - \sum \delta Q)$ reached a value of one; at that point, prices would be fully “rolled in.”

field prices, but under conditions of shortage, the regulated ceiling price P_c prevails rather than the market demand price P^* . This is illustrated in Figure 1. Under conditions of shortage, the quantity

FIGURE 1
SUPPLY AND DEMAND FOR ADDITIONAL RESERVES

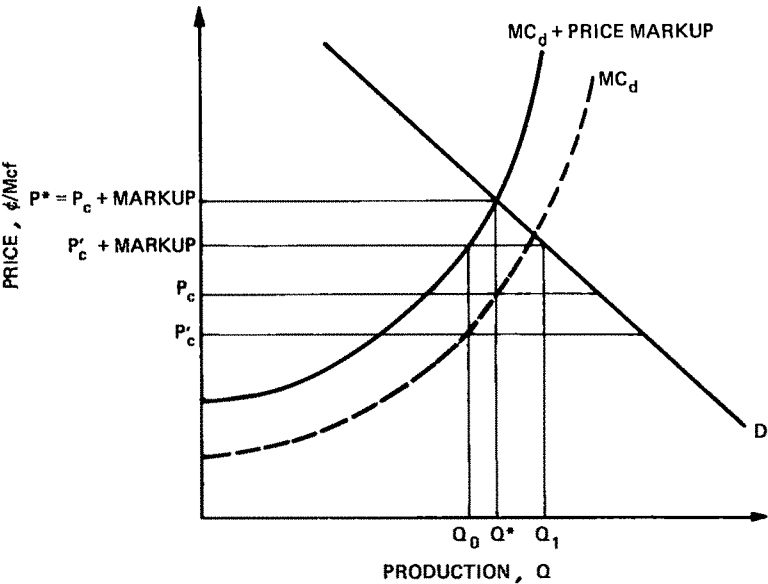


realized is on the supply but not on the demand function, with the result that the demand function is not observable. At the same time, production out of reserves is affected by the reserve shortage. The supply curve for gas production consists of a marginal development cost curve which represents the cost of incrementing gas production (by running existing gas development wells at higher capacity or by drilling new development wells). The demands for gas production consist of wholesale volumes taken by final buyers from the pipeline systems at prices equal to field prices plus the pipeline markups. The final demands could conceivably still be satisfied at the beginning of the reserve shortage, as a result of the pipelines' calling on existing reserves to produce at a higher rate.

Sufficient production under a condition of reserve depletion cannot be had indefinitely. Eventually, the amount of reserves available to back production is reduced, and supply of production at ceiling prices is reduced. As the reserve backing becomes smaller, production tends to fall, and a gap is opened between the demands for production and the supplies that will be made available.

The marginal cost curves, the price markup, and a wholesale demand curve are all shown in Figure 2. In this diagram, the ceiling price is sufficient to bring forth production Q^* which clears the market at wholesale price P^* , which is just the field price plus the pipeline's markup. Under these conditions, the demand curve for production is "registered in the market" or observable as a total quantity demand at one specific price level. However, if the regulated field price is reduced to a level P'_c , excess demand will result equal to $(Q_1 - Q_0)$. The supply of production is reduced by price disincentives to a level below the demands put on the pipeline system by retail gas utility companies serving final residential, industrial, and com-

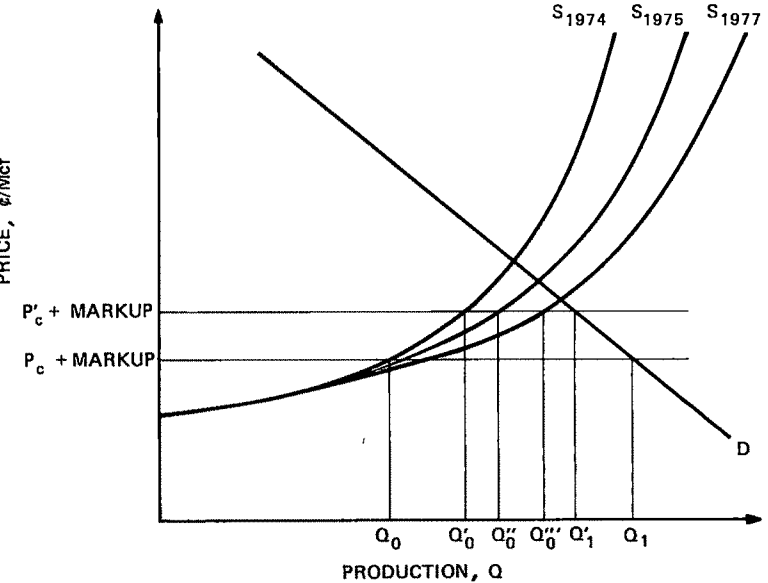
FIGURE 2
SUPPLY AND DEMAND FOR PRODUCTION



mercial consumers. There are shortages in both field reserve and final production markets.

How would an increase in the ceiling price eventually cause a reduction in the production shortage? Consider, for example, an increase in wellhead ceiling prices under FPC regulation. When the ceiling price is increased, the production cost curve remains fixed initially, since reserve levels would not change immediately, and therefore marginal production costs would not change. The higher price, however, would soon elicit more production out of given reserves Q_0' and also reduce demands for production from Q_1 to Q_1' (as shown in Figure 3). Thus, even within the short run, excess

FIGURE 3
SUPPLY AND DEMAND FOR PRODUCTION

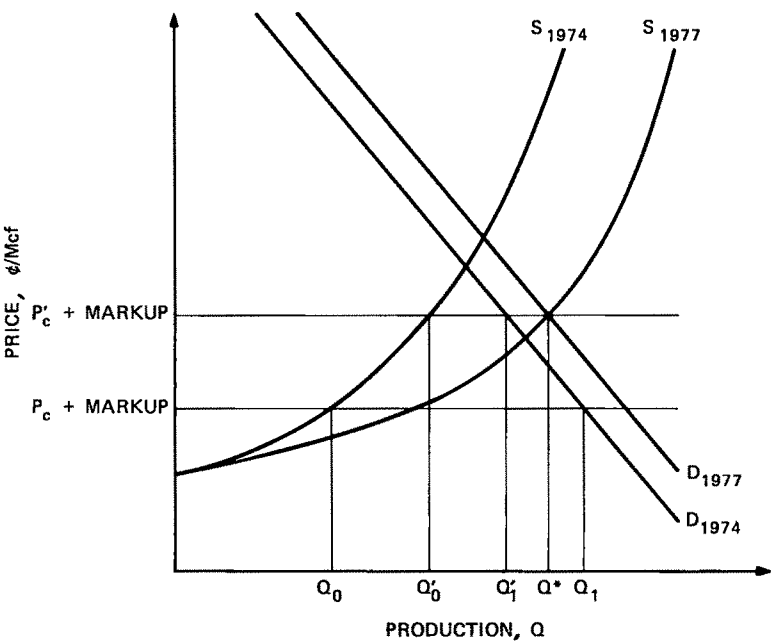


demand would be decreased by a price change from $(Q_1 - Q_0)$ to $(Q_1' - Q_0')$, although the change may not be substantial because the "roll-in" pricing practices of the Commission would dampen the demand effect.

After a few years, however, it is likely that reserve levels will have been increased as a result of the higher level of new contract prices. Higher ceiling prices would stimulate more exploratory drilling which, in turn, should result in new discoveries of gas that would add to reserve levels. At that point in time, a given level of production could be induced at lower marginal development costs because of the presence of higher reserve levels. The supply of production curve would have shifted to the right. Even if demand were to remain at Q_1' , an increase in supply to Q_0'' would reduce the extent of the shortage. After a few more years, the full effects of the ceiling price increase would have occurred, with the supply of production shifting farther to the right so that excess demand would fall to the even lower level of $(Q_1' - Q_0''')$.¹³

Of course, if we could increase the ceiling price by just the right amount—to shift the supply curve enough to make up for past excess demand and the present and independent shifts in the demand curve (stemming from increased population, national income, etc.)—we would reach a situation where there was no excess demand in 1977. This is shown in Figure 4. Note, however, that *until*

FIGURE 4
LONG-RUN EQUILIBRIUM



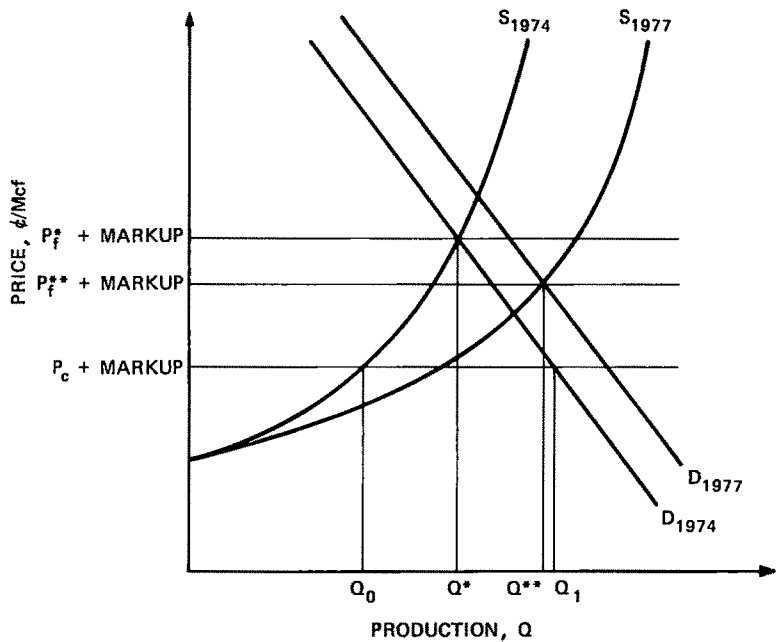
1977, there would still be some amount of excess demand. In the first year after the price increase, for example, excess demand will be

¹³ This analysis assumes that the pipeline markup is constant with respect to the level of production and that the difference between $(P_c + \text{price markup})$ and $(P_c' + \text{price markup})$ is equivalent to the "roll-in" price increase allowed under regulation. The econometric model discussed later deals with these matters of detail in specific price markup equations.

$(Q_1' - Q_0')$; not until 1977, when the supply curve, demand curve, and price line all intersect at the same point, would there be a level of production, Q^* , that resulted in no excess demand.

What if the field price of all gas were immediately and completely deregulated? This would result in a field price increase to the level P_f^* which, when the pipeline markup is added on, would clear the production market immediately (as shown in Figure 5; the supplies

FIGURE 5
EQUILIBRIUM UNDER COMPLETE DEREGULATION



and demands for production are both equal in one year to Q^*). This substantial price increase would not be the end of the story, however. After three or four years, the supply curve would have shifted to the right, again because increased exploration and discoveries in response to the price increase would have added substantially to the reserve base. The demand curve would perhaps also have shifted to the right after three or four years, but the net result would probably be a decrease in price and further increase in quantity of production over the four-year period. This is illustrated in Figure 5 where the 1974 equilibrium quantity Q^* is increased over time to Q^{**} as the equilibrium field price P_f^* falls to P_f^{**} .

These examples indicate that pricing policies have a three- or four-step effect upon the size of the shortage, and it may take several years before the full effects of a price change become apparent. The econometric model presented in the next section allows us to analyze these dynamic effects of alternative price changes quantitatively.

3. An econometric model for policy analysis

PAUL W. MACAVOY AND
466 / ROBERT S. PINDYCK

■ Most previous econometric studies of natural gas have investigated either demand or supply of gas, but have neglected the simultaneous interaction of these two sides of markets. Balestra, for example, in his classic study of the demand for natural gas by residential and commercial consumers, assumed a perfectly elastic

supply curve for production.¹³ This assumption was probably justified during the 1950s and 1960s, since production of gas for final consumers took place on an "as needed" basis from large stocks of reserves; but it would not continue to be valid during the 1970s, however, as total gas demand exceeds production constrained by smaller reserve levels. The supply studies of Erickson and Spann and of Khazzoom,¹⁴ similarly, are admirable attempts at defining and testing some of the relationships that exist in the gas industry, particularly those accounting for reduced reserve levels under price controls. But, to the extent that policies are changed in the future so that markets clear, and demand is once again observed, models of only the supply side of markets will be inadequate to represent the effects of policy. If the industry is to be properly understood, then, production and reserve supply-demand levels of the industry have to be analyzed as a simultaneous system.

The model developed as part of this study consists of a set of simultaneous econometric relationships among several policy-related variables. Variables endogenous to the field market include, on the supply side, nonassociated and (oil) associated discoveries of gas reserves, extensions and revisions of associated and nonassociated reserves, and wells drilled. These variables directly or indirectly depend on the field prices paid by pipelines in new contracts for gas. Field prices would be endogenous if demands could clear, but after ceiling prices were set by the FPC in the 1960s, this variable became an exogenous policy variable.

Endogenous variables in the wholesale market include demand for production of gas and wholesale prices for three wholesale delivery sectors: mainline industrial sales, sales for resale that are ultimately industrial, and sales for resale that are ultimately residential and commercial. Throughout the 1960s wholesale production demand was approximately satisfied, even though there was excess demand for reserves (of course, reserve-production ratios dropped dramatically during the decade), and thus wholesale demand equations can be estimated from data generated in this period.

An equation for marginal development costs (the "supply curve" for production in field markets), when combined with pipeline wholesale price markup equations, provides an estimate of volume of production sent to wholesale buyers. This allows us to determine excess demand by comparing estimated "production out of reserves" transmitted to the wholesale markets with estimated demands for production at wholesale.

There is no single field market, nor is there a single wholesale market, in the United States. Producers throughout the country do not take their gas to the Chicago Board of Trade in order to make offers of sale to pipelines. Rather, there are several "regional" field markets and several "regional" wholesale markets, and the natural gas industry is characterized by the spatial interrelationships of these markets. In our econometric model, field reserve and production equations are estimated for supply regions to account for the regionalization. Wholesale demand equations are estimated for each of five parts of the country (each part roughly representing a regional

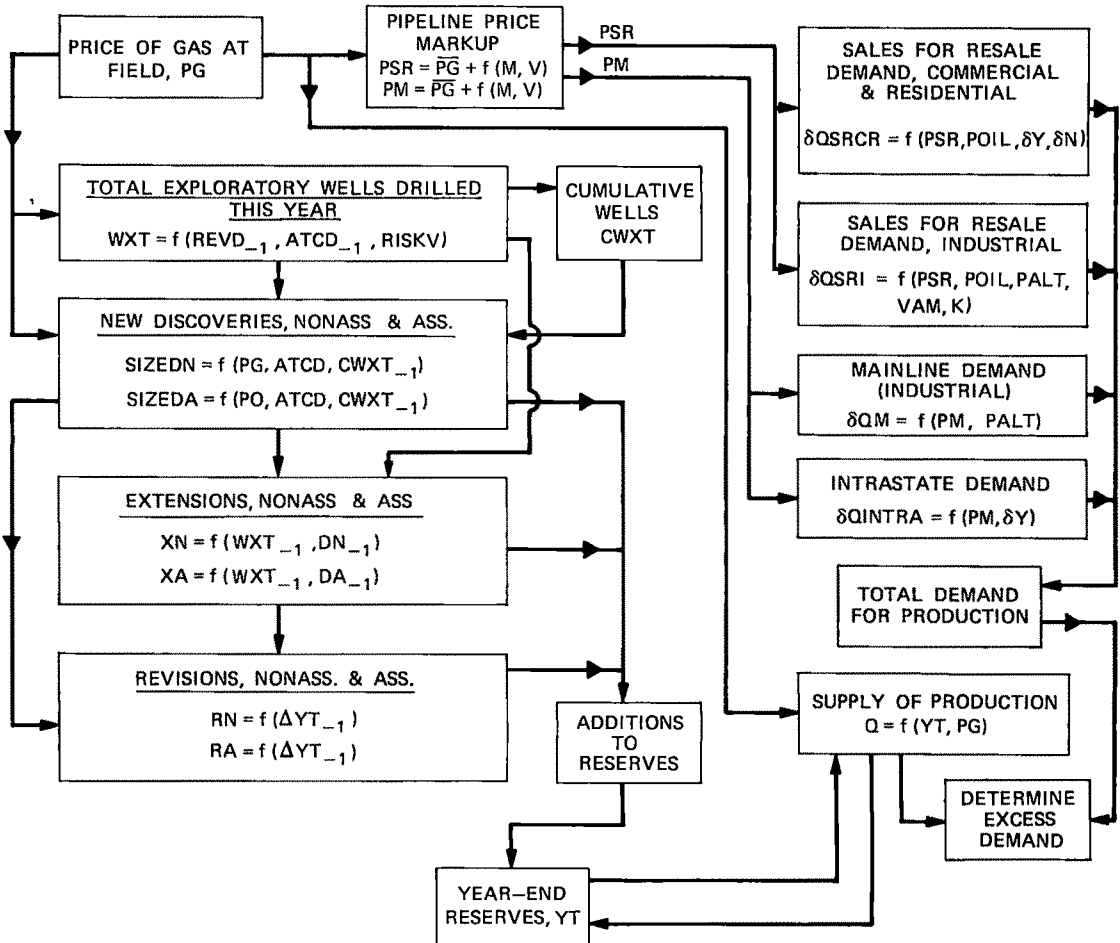
¹³ See [2].

¹⁴ See [8] and [14], respectively.

industrial and retail utility buyer's market). Gas from each production district in the country is allocated to one or more of the wholesale consumption regions, according to the average allocation proportions that prevailed in the past. In this way, excess demand can be computed on a regional and country-wide basis.

□ **Structure of the model.** The organization of the model is illustrated in Figure 6. Note, however, that this figure leaves out (for simplicity)

FIGURE 6
THE MODEL



the spatial interconnections between production districts and regional wholesale markets. In the model as it actually runs, the wholesale prices of gas (for mainline sales and for sales for resale) are computed for each region of the country by taking the wellhead price of gas at the production source and by adding a markup based on pipeline mileage and volumetric capacity. When a wholesale consumption region is supplied by more than one production district (as is usually the case), wellhead prices, mileages, and capacities are weighted according to previous actual proportions of production from each district. Let us now look at the individual parts of the model in more detail.

The field market

Probably the sector of the natural gas industry most difficult to capture in a conceptual model is the supply of new reserves. Most of the current controversy over regulatory policy centers on this sector—whether or not reserve additions have been “too low,” as a result of past regulatory policy. Since actual additions to reserves through new discoveries are realized by a complicated process involving a large number of technological factors, it may seem naive to try to model the process by using a set of simple econometric relationships. However, structural equations can be formulated to link economic and technological variables that are important in gas reserve additions and to describe most simply and directly the regulatory effects.

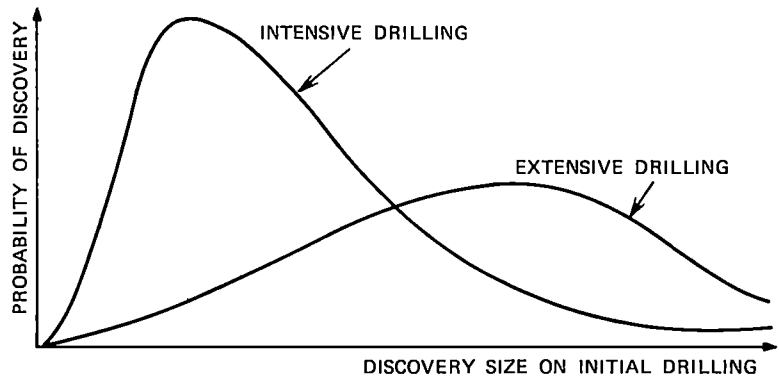
The major component of new reserve additions consists of new discoveries of both nonassociated and associated gas [nonassociated gas (N) is dry gas, while associated gas (A) includes both “dissolved” gas recovered from oil production and “free” natural gas forming a cap in contact with crude oil]. In our model, the discovery process begins with the drilling of wells, of which some will be successful in discovering gas, some will be successful in discovering oil (with associated gas), and some will be unsuccessful (i.e., dry holes). The drilling of wells depends largely on economic incentives. In our model, it is dependent upon prospective revenues from oil and gas production, average drilling costs, and a measure of drilling risk.

The model translates drilling activity into actual discoveries through two size-of-discovery variables—one for nonassociated gas and the other for associated gas. The size-of-discovery variable for nonassociated gas, for example, gives for any district and any year the average number of Mcf of gas discovered per well drilled. The size-of-discovery variables themselves are explained partly by economic variables (e.g., oil and gas prices and drilling costs) and partly by a depletion effect, in which extensive well drilling in the past (measured by cumulative wells drilled) makes it more difficult to discover gas in the present.

Drilling may be divided into two basic modes of behavior, depending on whether it is done *extensively* or *intensively*. In extensive drilling, few wells are drilled, but those that are drilled usually go out beyond the frontiers of recent discoveries to open up new geographical locations or previously neglected deeper strata at old locations. Typically, this would include drilling farther offshore, or onshore but at very great depth. Here the probability of discovering gas is rather small, but the size of discovery may be comparatively large. When drilling is done intensively, many wells are drilled in an area that has already proven itself to be a source for gas discovery. Here the probability of discovering gas is larger, but the size of discovery is likely to be very small. In Figure 7, typical probability distributions for discovery size are shown for each mode of drilling behavior. Relative to the intensive drilling mode, discovery size for extensive drilling has a larger expected value but also a larger variance.

The producer who is engaged in exploratory activity has, at any point in time, a choice as to whether any increases in his drilling activity will be extensive or intensive, and this choice will be influenced by changes (or expected changes) in economic variables

FIGURE 7
DISCOVERY DISTRIBUTION



—field prices and drilling costs.¹⁵ The actual influence of these variables will depend on the producer's geological portfolio (i.e., the set of regions over which he has drilling rights), as well as on his translation of present and past prices and costs into expectations of future prices and costs. Suppose, for example, that drilling costs decrease. As a result, a producer might decide to accept greater risk and drill on the extensive margin, with the result that average discovery size will (*ex post*) increase. On the other hand, the producer may own partially drilled reservoirs that now are worth drilling out. If this is the case, he might decide to drill on the intensive margin with a resulting *ex post* decrease in average discovery size. Thus it is not possible *a priori* to determine whether the effect of a decrease in drilling costs on average discovery size will be positive or negative. The same is true for the effects of a change in the price of gas.

Higher gas prices should indeed result in more drilling, but we would expect that over the years the size of finds will decrease as our finite resource stock begins to get depleted. One may model the exploration and discovery process stochastically as sampling without replacement, so that the expected value of discovery size would decrease as the sampling process went on.¹⁶ It is not our object to try to predict how big the total stock of gas yet to be discovered is,¹⁷ but we would like to embody a "depletion effect" in our model, at least to the point of being able to extrapolate the long-run decreasing returns to industry size that have occurred over the past decade. We do this by including cumulative wells drilled as an explanatory variable in our "size-of-discovery" equations. If the level of drilling activity is the same next year as it is this year, we would expect to see the level of new discoveries drop somewhat, and this is what would indeed happen if discovery size depended negatively on cumulative wells drilled.

Additions to gas reserves can also occur as a result of extensions and revisions of existing fields and pools. These extensions and re-

¹⁵ See Fisher [9].

¹⁶ See Kaufman [13].

¹⁷ Industry spokesmen have claimed that higher gas prices result not only in more drilling activity, but also in a shift to the extensive mode resulting in larger discovery size. While we would expect higher gas prices to elicit more drilling, it is not clear that the additional drilling will be more extensive. Our results do show a positive relationship between price and discovery size, but with a low elasticity.

visions could be expected to depend on (1) price incentives, (2) past discoveries of gas, (3) existing reserve levels, and (4) the cumulative effect of past drilling. Extensions, which are influenced by (1), (2), and total drilling activity, are somewhat easier to model than revisions.¹⁸ Revisions of established reserve levels, on the other hand, are often erratic and difficult to predict. Although there may be some effect from prices in certain regions, revisions are basically proportional to prior discoveries and reserve levels.

New discoveries (DN , DA), extensions (XN , XA), and revisions (RN , RA) are combined to form additions to reserves. Aside from changes in underground storage (ΔUS), which we model as a constant percentage of production, the only major subtraction from reserves occurs as a result of production (Q). Thus, for any time t , in a production district j , total gas reserves are given by the identity:

$$R_{t,j} = R_{t-1,j} + DN_{t,j} + XN_{t,j} + RN_{t,j} + DA_{t,j} + XA_{t,j} + RA_{t,j} - Q_{t,j} - \Delta US_{t,j}. \quad (1)$$

Since the supply of new reserves is so determined, if the wellhead price of gas were not regulated, or if regulation were ineffective, then the demand for new reserves could be given by an equation for pipeline offers to buy reserve commitments at specified new contract wellhead prices. Since 1962, however, there has been excess demand for new reserves, and thus the demand function for new reserves has not been observable. Instead, the price has been given by the exogenous ceiling price.

Production out of reserves

The supply of production as a function of price is simply the marginal cost (in the short term) of developing existing reserves so that a particular level of annual flow can be achieved (e.g., by drilling development wells and then operating them). Clearly, marginal production costs will depend on reserve levels relative to production, and as the reserve-to-production ratio becomes small, we would expect marginal costs to rise sharply.

Let us examine what marginal costs would be corresponding to a production level q out of proved reserves R .¹⁹ Assuming a constant decline rate, a , in percent per year,

$$a = q/R = 1/\text{reserve-production ratio}, \quad (2)$$

we can write the proved reserve level as

$$R = q \int_0^{\infty} e^{-at} dt = q/a. \quad (3)$$

Then for a discount rate δ the "present-Mcf-equivalent" (PME) of a constant production level q is:

$$PME = q \int_0^{\infty} e^{-(a+\delta)t} dt = q/(a + \delta). \quad (4)$$

¹⁸ Extensions can result from either exploratory or development well drilling. Our model does not explain development well drilling, and therefore only exploratory wells are used to explain extensions.

¹⁹ Our thanks to M. Adelman and M. Baughman for their assistance on this part of the model.

Now we assume that the development investment, I , needed to obtain the production level q is given by:

$$I = A + ce^{\beta a} q, \quad (5)$$

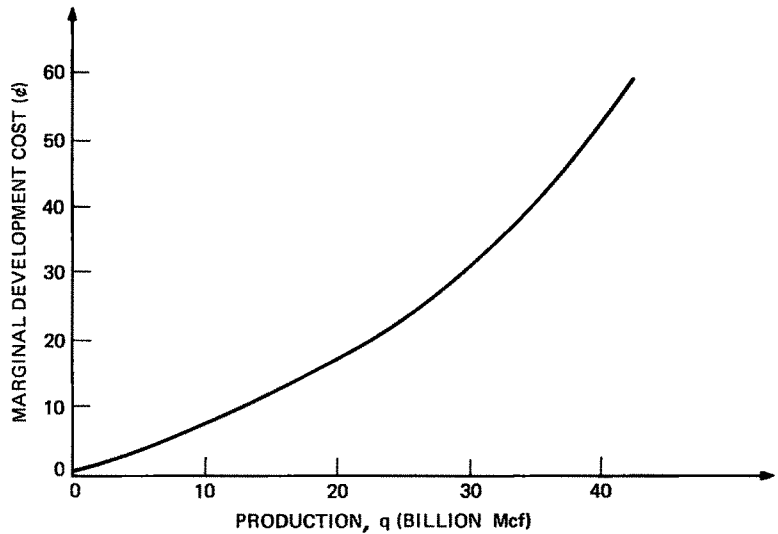
where A is a start-up cost, c is constant over the range of zero well interference, and β is a parameter with value around 10. Thus, when a is small (e.g., the reserve-production ratio is larger than 10), I will be roughly linear in q , but when a becomes larger (e.g., the reserve-production ratio approaches 5), an exponential rise in costs begins to predominate. The marginal development cost (MDC) is then given by:

$$\begin{aligned} MDC &= \frac{dI}{d(PME)} = \frac{dI}{dq} \cdot \frac{dq}{d(PME)} \\ &= \left(\frac{\partial I}{\partial a} \frac{da}{dq} + \frac{\partial I}{\partial q} \right) \cdot \frac{dq}{d(PME)} \end{aligned} \quad (6)$$

$$\begin{aligned} MDC &= \left(\frac{c\beta}{R} e^{\beta a} q + ce^{\beta a} \right) \cdot \frac{(a + \delta)^2}{\delta} \\ &= (\beta a + 1) ce^{\beta a} \frac{(a + \delta)^2}{\delta} \\ &= (\beta a + 1) c \delta e^{\beta a} \left(1 + \frac{a}{\delta} \right)^2. \end{aligned} \quad (7)$$

This marginal cost curve is illustrated in Figure 8 for $\delta = 0.1$, $\beta = 10$, $c = 10$, and $R = 0.2$ trillion Mcf . For small values of a

FIGURE 8
HYPOTHETICAL MARGINAL COST CURVE



(i.e., large reserve-production ratios), the curve is predominantly quadratic, and when a becomes large, the curve begins to look more like an exponential. We estimate a marginal cost curve (which becomes our supply curve for production out of reserves when price is set equal to marginal cost) that is essentially exponential. Aside from

the fact that this gives a better fit to recent data, it is also in keeping with our goal of calculating excess demand for gas under conditions of declining reserve-production ratios.

The wholesale market

Supply of production out of reserves has to be measured against demands for that production after it has been transmitted to wholesale markets by pipelines. The demands for production are approximated by curves fitted on a disaggregated basis, namely by wholesale demand equations for (1) gas sales for resale (split into commercial-residential gas and industrial gas, on the basis of percentages distributed to these two groups for ultimate consumption), (2) gas sales directly off the pipelines for consumption, and (3) intrastate sales by producers and pipelines to final consumers. The wholesale prices of gas (disaggregated into a "sales-for-resale" price and a "mainline sales" price) are computed by adding a markup to the field price based on (1) the mileage between the production district and the consuming region, and (2) the volumetric capacity of the pipelines.

The demand equations follow a general formulation, in which the quantity demanded is dependent on wholesale price, the price of alternative fuels, and "market size" variables, such as population, income, and investment, which determine the number of potential consumers. In all of the equations, the dependent variable will be *new demand*, δQ , rather than the level of total demand. In the short run, as Balestra has shown for residential gas,²⁰ the level of total demand should be relatively price inelastic and would depend on stock variables that do not change much in time (e.g., the total stock of gas-burning appliances for residential gas). New demand, however, should respond to the price of gas and to the price of competing fuels (decisions to buy new appliances, for example, are affected by fuel prices). The new demand for gas, δQ , is made up of the increment in gas consumption $\Delta Q = Q_t - Q_{t-1}$, and of replacement for continuation of old consumption. To find replacement, total residential and commercial gas demand could be considered to be a function of the stock of gas-burning appliances, A :

$$Q_t = \lambda \cdot A_t, \quad (8)$$

where λ is the (constant) utilization rate. Then, if r is the average rate at which the stock of appliances depreciates, the replacement demand for gas includes $r\lambda A_{t-1}$, and total new demand is

$$\delta Q_t = \Delta Q_t + r\lambda A_{t-1}. \quad (9)$$

Now substituting (8) into (9) gives:

$$\delta Q_t = \Delta Q_t + rQ_{t-1}. \quad (10)$$

Thus, new demand for gas is the sum of the incremental change in total gas consumption (ΔQ_t) plus the demand resulting from the replacement of old appliances.

Our *a priori* assumption is that *new* demand depends on prices and total income (through purchases of new appliances), and that the *level* of *total* demand is itself a function of income and popula-

²⁰ In [2].

tion. Thus, we have for residential and commercial demand:

$$\delta QSR_{t,k} = f(PSR_{t,k}, POIL_{t,k}, \delta Y_{t,k}, \delta N_{t,k}, K_{t,k}), \quad (11)$$

where PSR is the sales-for-resale wholesale price, $POIL$ the average price of oil, Y disposable personal income, N the population by state, and K new capital expenditures, all in region k , and

$$\delta Y_{t,k} = \Delta Y_{t,k} + rY_{t-1,k} \quad (12a)$$

and

$$\delta N_{t,k} = \Delta N_{t,k} + rN_{t-1,k}.^{21} \quad (12b)$$

The form for industrial demand is similar to that for residential and commercial demand. The prices of gas and oil at wholesale determine additional demands, and the "size of the market," as measured by value added in manufacturing or capital outlay, also sets new demands. The equation is:

$$\delta QSRI_{t,k} = f(PSR_{t,k}, POIL_{t,k}, VAM_{t,k}, K_{t,k}). \quad (13)$$

The model is closed by spatially interconnecting production districts with consuming regions. A flow network is constructed which, based on relative flows calculated from 1970 data, determines where each consuming region obtains its gas. Average wholesale prices (again both for sales for resale and for mainline sales) can thus be computed for each consumption region in the country, since mileages and volumetric capacities are then determined. The equation form is:

$$PSR_{t,k} = \overline{PG}_{t,k} + f(M_{t,k}, V_{t,k}) \quad (14)$$

for "rolled-in" average field price $\overline{PG}$, mileage M , and volumetric pipeline capacity, V . Given the markup, wholesale demand (by type) is then also computed for each region of the country. Wholesale demands can be summed to produce total demand for each region of the country, and since we know what the transmission of production will be to each region of the country, we can determine excess demand.

□ **Estimation of the model.** The model was estimated using pooled cross-section and time-series data. The time bounds of the regressions are different for different equations, not only because of data limita-

²¹ Balestra [2] distinguishes between two depreciation rates, one for gas appliances and the other for alternative fuel-burning appliances, since lifetimes for appliances using alternative fuels may be different. He then estimates the two depreciation rates by estimating an equation of the form:

$$QSR_{t,i} = a_0 + a_1 PSR_t + a_2 \Delta N_t + a_3 N_{t-1} + a_4 \Delta Y_{t-1} + a_5 Y_{t-1} + a_6 QSR_{t-1}. \quad (i)$$

The depreciation rate for gas appliances is then given by $(1 - a_6)$. (His results, however, gave an estimated value of a_6 that was always greater than one, which cannot be justified theoretically.) The all-fuel depreciation rate comes out of equation (i) as either the ratio a_3/a_2 or a_5/a_4 . Thus, the equation is overidentified, and the depreciation rate can be obtained only by estimating (i) subject to the constraint of $a_3/a_2 = a_5/a_4$. (The resulting estimation problem is nonlinear, but Balestra uses an iterative method suggested by Houthakker and Taylor [12] to obtain an estimated depreciation rate equal to 0.11.) Following Balestra in estimating these depreciation rates, we obtained a single rate for the West region equal to 0.07 and use this both for industrial and for residential and commercial demand.

tions, but also because of structural changes over time in the industry. Wholesale demand equations, for example, were estimated using data only from 1964 to 1970, even though data were available from as far back as 1960, because it was felt that demand elasticities had changed considerably during the 1960s as a result of new air pollution standards.²²

Cross sections were also different for different equations. Field market equations were estimated by pooling data from all of 18 FPC production districts, while individual sales-for-resale wholesale demand equations were estimated over what were considered to be the proper regional wholesale markets, and thus each used data pooled from five to ten states. District breakdowns and time bounds are summarized for all equations of the model in Table 1.

TABLE 1
CROSS SECTIONS AND TIME BOUNDS FOR THE MODEL'S STOCHASTIC EQUATIONS

EQUATIONS	DISTRICTS POOLED	TIME BOUNDS
WELLS (WXT)	ALL 18 FPC PRODUCTION DISTRICTS*	1964 -1971
DISCOVERY SIZE (SIZEDN, SIZEDA)	ALL 18 FPC PRODUCTION DISTRICTS	1964 -1971
EXTENSIONS & REVISIONS (XN, XA, RN, RA)	ALL 18 FPC PRODUCTION DISTRICTS	1964 -1971
PRODUCTION FROM RESERVES (QS)	ALL 18 FPC DISTRICTS	1962 -1971
PRICE MARKUPS (PSR, PM)	40 SALES-FOR-REALE REGIONS (SEE BELOW), PLUS 15 MAINLINE SALES REGIONS (SEE BELOW)	1956 -1969
SALES-FOR-REALE DEMAND (RESIDENTIAL/COMMERCIAL AND INDUSTRIAL)		
NORTHEAST	NEW ENGLAND, NEW JERSEY, NEW YORK, PENNSYLVANIA, OHIO, MARYLAND + DELAWARE + WASHINGTON, D C., VIRGINIA, WEST VIRGINIA	1964 - 1970
NORTH CENTRAL	ILLINOIS, INDIANA, MICHIGAN, WISCONSIN, IOWA, MINNESOTA, MISSOURI, NEBRASKA, SOUTH DAKOTA	1964 - 1970
WEST	ARIZONA, COLORADO, IDAHO, NEVADA, NEW MEXICO, UTAH, WYOMING, CALIFORNIA, WASHINGTON, OREGON	1964 - 1970
SOUTH CENTRAL	KANSAS, ARKANSAS, OKLAHOMA, TEXAS, MISSISSIPPI, LOUISIANA	1964 - 1970
SOUTHEAST	FLORIDA, GEORGIA, NORTH CAROLINA, SOUTH CAROLINA, ALABAMA, KENTUCKY, TENNESSEE	1964 - 1970
MAINLINE DEMAND (INDUSTRIAL)		
GAS-PRODUCING STATES	COLORADO + WYOMING, KANSAS, KENTUCKY + OHIO + PENNSYLVANIA + WEST VIRGINIA, LOUISIANA, MISSISSIPPI, OKLAHOMA + TEXAS	1964 - 1970
NONPRODUCING STATES	ALABAMA + GEORGIA, ARIZONA, ARKANSAS, FLORIDA, ILLINOIS, IOWA + NEBRASKA, MINNESOTA, MISSOURI, TENNESSEE	1964 - 1970
INTRASTATE SALES	CALIFORNIA, NEW MEXICO, WEST VIRGINIA, KANSAS, TEXAS, OKLAHOMA, LOUISIANA	1964 - 1970
*THESE INCLUDE TEXAS 1, 2, 3, 4, 6, 9, 10, CALIFORNIA, COLORADO-UTAH, KANSAS, LOUISIANA NORTH, LOUISIANA SOUTH, MISSISSIPPI, NEW MEXICO NORTH, PERMIAN (=NEW MEXICO SOUTH + TEXAS 7C + TEXAS 8 + TEXAS 8A), OKLAHOMA, WEST VIRGINIA + KENTUCKY, WYOMING		

Statistical results

The regression results described below were obtained by using two-stage least squares²³ whenever unlagged endogenous variables

²² A test of this hypothesis and a more detailed study of demand will appear in a future paper.

²³ It should be pointed out that this model is "almost" block recursive, i.e., it can be broken up into two large blocks of equations (one determining reserves and the other determining production), and the simultaneous interaction of these

appeared on the right-hand side of an equation. All of the equations are linear in form, with the exception of the equation describing production out of reserves, which is logarithmic in the price term (marginal production costs are an exponential function of production and reserves). The *t*-statistics are shown in parentheses below each estimated coefficient. Also listed for each equation are the R^2 , *F*-statistic, standard error of the equation, Durbin-Watson statistic, and the mean of the dependent variable. Note that the Durbin-Watson has limited meaning, since error terms may be autocorrelated across time and/or across cross sections, and these effects are not separated.²⁴

One of the difficulties in constructing a model of this sort is that one must work under the constraints imposed by the previously mentioned data limitations. Data for many variables are either difficult or else impossible to obtain, particularly for years prior to 1966. In addition, many of the data are extremely noisy. As a result, a good deal of compromise was often required in estimating equations between functional forms that were theoretically pleasing and those that lent themselves to the existing data. This should be kept in mind when interpreting the estimation results.

Field market equations

The field market portion of the model contains seven stochastic equations that explain total exploratory well drilling (*WXT*), nonassociated and associated average discovery size (*SIZEDN*, *SIZEDA*), extensions (*XN*, *XA*), and revisions (*RN*, *RA*). Non-associated and associated new discoveries (*DN*, *DA*) can be determined from the two identities:

$$DN_{t,j} = SIZEDN_{t,j} \cdot WXT_{t,j} \quad (15)$$

$$DA_{t,j} = SIZEDA_{t,j} \cdot WXT_{t,j} \quad (16)$$

and the supply of new reserves is then determined from the identity in equation (1).

Exploratory well drilling responds to three economic incentives, all of which are exogenous to the model. The first of these, a surrogate for anticipated returns from exploration, is total revenues (deflated by a GNP price index), *REVD*, from sales of new oil and gas at the wellhead. Exploratory drilling may result in the discovery of either gas or oil, so total revenues are used as an explanatory variable. Note that changes in the price of gas (or the price of oil) can affect drilling activity through this revenue variable.

Average total drilling costs (also in deflated terms), *ATCD*, are a second explanatory variable; rising costs are expected to have a

blocks is weak. For example, equations (17) through (23), together with the exogenous ceiling wellhead price, determine equilibrium in the market for new reserves, and while additions to reserves certainly have a strong impact on wholesale production markets, this impact occurs over a number of years since price increases are rolled in and marginal production depends on total reserves and not new reserves. Thus, two-stage least squares was applied on a block-by-block basis.

²⁴ It is reasonable to expect, in fact, that the error terms are autocorrelated both across time and across districts. We did not attempt to correct for the problem of serial correlation in this version of the model (although we shall in a future version). District heterogeneity can be one source of cross-sectional autocorrelation, and this was dealt with at least in part by the use of district dummy variables.

negative impact on drilling. The third explanatory variable, *RISKV*, provides a measure of relative risk between different regions. It is a purely cross-sectional variable that does not change in time, and is the sample variance (measured over recent years) of payoff size in each district. Finally, three dummy variables are introduced (*DDA*, *DDB*, *DDC*) to account for heterogeneity among four broadly defined regional field markets in the United States. The estimated equation is:

$$\begin{aligned}
 WXT = & 439.39 + 934.53 DDA + 83.75 DDB + 137.48 DDC \\
 & (9.17) \quad (3.81) \quad (0.40) \quad (3.24) \\
 & + 3.156 \times 10^{-4} REVD_{t-1} - 4.43 \times 10^{-3} ATCD_{t-1} \\
 & \quad (2.22) \quad (-4.79) \\
 & \quad \quad - 2.13 \times 10^{-7} RISKV \\
 & \quad \quad \quad (-2.74) \quad (17)
 \end{aligned}$$

$$R^2 = 0.495 \quad F = 17.2 \quad S.E. = 193.6 \quad DW = 0.36$$

$$\text{Mean}(WXT) = 359.2.$$

Here it can be seen that "cash flow" (as measured by *REVD*_{*t-1*}) has a positive effect on drilling, while costs and risk have negative effects. All three effects are statistically significant. Thus, drilling increases as lagged prices increase and as lagged costs and risk decrease.

Economic variables influencing the size of nonassociated discoveries per well include the wellhead price of gas, *PG*, and average drilling costs per well of exploratory drilling, *ATCD*. As explained earlier, the signs of these variables cannot be predicted *a priori*. A third explanatory variable is the cumulative number of wells drilled, *CWXT*. This is expected to have a negative impact on discovery size, since it represents a depletion effect. Finally, three district dummy variables are added to the equation. The resulting equation is:

$$\begin{aligned}
 SIZEDN = & 815.62 - 600.69 DDA + 189.70 DDB + 825.43 DDC \\
 & (-2.78) \quad (-0.94) \quad (0.63) \quad (6.52) \\
 & + 28.89(PG_{t-1} + PG_{t-2} + PG_{t-3})/3 \\
 & \quad (2.05) \\
 & + 0.0131(ATCD_{t-1} + ATCD_{t-2} + ATCD_{t-3})/3 \\
 & \quad (5.86) \\
 & \quad \quad - 0.301 CWXT_{t-1} \\
 & \quad \quad \quad (-3.18) \quad (18)
 \end{aligned}$$

$$R^2 = 0.631 \quad F = 29.9 \quad S.E. = 578.3 \quad DW = 0.74$$

$$\text{Mean}(SIZEDN) = 602.4.$$

The equation explaining the size of associated discoveries has the same form, except that the price of oil, *PO*, is used instead of the price of gas:

$$\begin{aligned}
 SIZEDA = & 50.60 + 106.96 DDA + 48.34 DDB + 30.02 DDC \\
 & (1.64) \quad (2.65) \quad (2.67) \quad (4.02) \\
 & - 14.30(PO_{t-1} + PO_{t-2} + PO_{t-3})/3 \\
 & \quad (-1.64) \\
 & + 3.138 \times 10^{-4}(ATCD_{t-1} + ATCD_{t-2} + ATCD_{t-3})/3 \\
 & \quad (2.24) \\
 & \quad \quad - 4.590 \times 10^{-3} CWXT_{t-1} \\
 & \quad \quad \quad (-1.84) \quad (19)
 \end{aligned}$$

$$R^2 = 0.601 \quad F = 26.3 \quad S.E. = 34.6 \quad DW = 1.06$$

$$\text{Mean}(SIZEDA) = 38.5.$$

Together, these two reserves equations show strong lag effects by price and cost averaging and strong depletion effects (with the coefficients of most variables statistically significant). The price effects are in opposite directions. The price of gas has a strong positive effect on size of nonassociated discoveries while the price of oil has a negative (but insignificant) effect on the size of associated discoveries. These effects can occur in this combination because of aggregation of intensive and extensive drilling across districts; their net impact, as prices increase, however, is likely to be increased discoveries.

Extensions of nonassociated gas depend on previous gas discoveries (relatively higher gas finds are a stimulus to extend gas rather than oil fields), and on total exploratory drilling. Exploratory drilling is used as a proxy for total exploratory *and* development drilling. (Extensions also result from the drilling of development wells, but this variable is not included in the model.) The equation also includes district dummy variables to account for the significant cross-sectional heterogeneity:

$$\begin{aligned} XN = & 2.204 \times 10^4 + 1.33 \times 10^6 DDA + 1.56 \times 10^6 DDB \\ & (0.430) \quad (5.63) \quad (11.04) \\ & + 3.639 \times 10^5 DDC + 237.1 WXT_{t-1} + 0.137 DN_{t-1} \\ & (0.55) \quad (1.93) \quad (1.24) \end{aligned} \quad (20)$$

$$R^2 = 0.725 \quad F = 67.6 \quad S.E. = 3.2 \times 10^6 \quad DW = 0.87$$

$$\text{Mean}(XN) = 2.88 \times 10^6.$$

The equation for extensions of associated gas is altogether comparable in form; associated discoveries replace nonassociated discoveries, but otherwise the dummy regional variables and the wells variable are the same:

$$\begin{aligned} XA = & -1054.93 + 2.323 \times 10^6 DDA + 1.205 \times 10^5 DDB \\ & (-0.11) \quad (5.20) \quad (4.44) \\ & - 1947.44 DDC + 48.00 WXT_{t-1} + 0.722 DA_{t-1} \\ & (0.17) \quad (2.11) \quad (2.53) \end{aligned} \quad (21)$$

$$R^2 = 0.693 \quad F = 57.7 \quad S.E. = 5.8 \times 10^4 \quad DW = 1.60$$

$$\text{Mean}(XA) = 4.43 \times 10^4.$$

Revisions of nonassociated gas (*RN*) depend on the same regional dummies and on the *change* in the previous year's reserves of gas. Large short-run increases in reserves increase the revisions in the following year (which in fact are adjustments to established discoveries). The equation is as follows:

$$\begin{aligned} RN = & 6.87 \times 10^4 + 1.138 \times 10^6 DDA - 6.864 \times 10^6 DDB \\ & (1.19) \quad (5.95) \quad (-3.60) \\ & + 3.438 \times 10^4 DDC + 0.239 \Delta YT_{t-1} \\ & (0.373) \quad (4.963) \end{aligned} \quad (22)$$

$$R^2 = 0.398 \quad F = 18.2 \quad S.E. = 4.5 \times 10^5 \quad DW = 1.65$$

$$\text{Mean}(RN) = 2.66 \times 10^4.$$

Note that there is no long run effect from the ΔYT_{t-1} term. Thus, a one Mcf increase in the *stock* of reserves implies about 0.24 Mcf increase in the *flow* of nonassociated revisions one year later.

Associated revisions (RA) are extremely erratic and difficult to explain with a simple linear regression model. There seems to be no relationship between this variable and lagged reserves of gas, primarily because associated revisions are linked more closely to oil reserves (which are not explained in this model). We modeled associated revisions by relating them to the previous year's change in reserves and the regional dummies, although the relationship is somewhat dubious:

$$RA = -2234.90 + 4.863 \times 10^5 DDA + 3.240 \times 10^5 DDB \\ (-0.07) \quad (4.56) \quad (3.04) \\ + 3.550 \times 10^4 DDC + 0.106 \Delta YT_{t-1} \quad (23) \\ (0.69) \quad (3.94)$$

$$R^2 = 0.289 \quad F = 11.2 \quad S.E. = 2.5 \times 10^6 \quad DW = 1.71 \\ \text{Mean}(RA) = 1.83 \times 10^4.$$

Production out of reserves

Two equations are estimated to predict the supply of production out of reserves, one for Louisiana South, and another for the remainder of the United States. Louisiana South is a large producing region with cost characteristics somewhat different from the rest of the country, and is therefore treated separately. Although data were available from 1955 to 1971, the nationwide production equation was estimated using only data from 1965 to 1971, since it was only in these years that production approached full short-term capacity. The Louisiana South equation, however, is estimated over the range 1962 to 1971 in order to have a sufficient number of degrees of freedom (only one production district is involved, so there are only as many data points as there are years).

In both equations, total production Q is regressed against the log of the wellhead price PG and against total reserves YT . This yields a supply of production equation of the form:

$$PG_{t,j} = \alpha_0 e^{\alpha_1 Q_{t,j} - \alpha_2 YT_{t,j}}, \quad (24)$$

with α_0 , α_1 , and α_2 all positive. This approximates (for reserve-production ratios less than 10) the marginal cost curve of equation (7). The nationwide equation also contains two regional dummy variables, one for the Permian region (DDB) and the other for the region including Oklahoma, Kansas, and Texas Railroad Commission Districts 2, 3, and 4 (DDC). These dummy variables help to account for district heterogeneity that remains even after Louisiana South is excluded from the sample.

The estimated equations are shown below.

United States, excluding Louisiana South (1965 to 1971):

$$Q = -5.053 \times 10^6 + 1.823 \times 10^6 DDB + 5.879 \times 10^6 DDC \\ (-1.30) \quad (16.10) \quad (7.43) \\ + 2.4755 \times 10^6 \log(PG) + 0.0240 YT \quad (25) \\ (1.85) \quad (5.57)$$

$$R^2 = 0.838 \quad F = 122.5 \quad S.E. = 2.59 \times 10^6 \\ \text{Mean}(Q)^{25} = 8.117 \times 10^5.$$

²⁵ This mean pertains to an average district, since we are pooling cross-section and time series data.

Louisiana South (1962 to 1971):

$$Q = -4.663 \times 10^7 + 1.0465 \times 10^7 \log(PG) + 0.2576 YT \quad (26)$$

(-11.25) (9.16) (8.45)

$$R^2 = 0.964 \quad F = 79.7 \quad S.E. = 4.53 \times 10^6$$

$$\text{Mean}(Q) = 4.941 \times 10^6.$$

Both of these equations show positive and significant effects of prices and total reserves. Thus, with higher prices, both short- and long-run production should increase (as in Figures 5 and 8).

Wholesale price markups

We assume that as long as pipelines are operating within their capacity, the markups over field prices are a function of mileage M_k and volumetric capacity V_k of the lines transmitting to region k . Without regulation, the pipelines would set markups so that the sales-for-resale price would be

$$PSR_{t,k} = [e/(e+1)]\overline{PG}_{t,k} + [e/(e+1)]MC_{t,k},$$

where e is demand elasticity in region k , $\overline{PG}$ is field price, and $MC_{t,k}$ is marginal costs of transmission to k . With regulation, markups on the field prices are prohibited [or $e/(e+1) = 1$] and the markups on $MC_{t,k}$ might be reduced. Transmission costs $MC_{t,k}$ are a function of M_k , $V_{t,k}$. The wholesale price equations thus have the simple form:

$$PSR_{t,k} - \overline{PG}_{t,k} = f(M_{t,k}, V_{t,k}) \quad (27)$$

for price on "sales for resale" to retail gas utilities and

$$PM_{t,k} - \overline{PG}_{t,k} = f(M_{t,k}, V_{t,k}) \quad (28)$$

for price on "mainline" sales to industrial buyers. Here, $\overline{PG}_{t,k}$, the "roll-in" wellhead price for wholesale region k , is determined by averaging the wellhead prices from flowing production in those regions which feed pipelines that transmit gas to region k . The mileage variable $M_{t,k}$ and the capacity variable $V_{t,k}$ are similarly defined, i.e., in terms of the average values of these variables from production regions to the wholesale market region k . The "markup" is the excess of the price over field costs as calculated by the coefficients of the independent variables in mileage and capacity.

Both of the explanatory variables M and V have been defined for each state with respect to five final sales regions. Regression equations have been calculated for each region; since they do not differ significantly from the national equation, only the national is reported. For sales for resale, the wholesale price is given by:

$$PSR = \overline{PG} + 12.561 + 0.712M - 3.3185 \times 10^{-4}V$$

(18.28) (18.66) (-10.57) (29)

$$R^2 = 0.561 \quad F = 356.2 \quad S.E. = 5.823 \quad DW = 2.25$$

$$\text{Mean}(PSR) = 17.42.$$

Note that the coefficient of volumetric capacity V is negative, since a larger capacity implies lower average costs. For mainline sales, our markup equation is given by:

$$PM = \overline{PG} + 5.365 + [0.914 + 0.446DDA]M - 5.895 \times 10^{-5}V \quad (30)$$

(4.73) (12.29) (5.96) (-1.76)

$$R^2 = 0.825 \quad F = 158.8 \quad S.E. = 2.45 \quad D.W. = 2.16$$

$$\text{Mean}(PM) = 13.91.$$

Here the mileage variable has a district dummy because of significantly higher markups in the Middle Atlantic states.

Wholesale demand for gas

Wholesale demand is broken into three major categories: sales-for-resale demand, mainline sales demand, and intrastate demand. Sales-for-resale demand is further broken down into gas that ultimately is resold for residential/commercial consumption and gas for industrial consumption, and for each category separate equations are estimated for each of five regions of the country. Mainline sales were broken down into a number of classifications. Attempts were made to calculate regression equations separately for mainline demand in producing states and for mainline demand in nonproducing states, on the hypothesis that the first set of demands would be more price elastic (since buyers could go to intrastate producers). With the data available, however, the differences were not observable and these categories were dropped. Intrastate sales were split between those in the South Central region (which were included in the total South Central equation) and those elsewhere.

New or additional demand $\delta Q = \Delta Q + rQ_{t-1}$ is used as the dependent variable in all equations. Assuming that all fuel-burning equipment has an average lifetime of 14 years, we have chosen a single depreciation rate r equal to 0.07. Explanatory variables besides the price of gas include the wholesale price of oil ($POIL$), income (δY), population (δN), value added in manufacturing (VAM), capital investment by industry (K), and a price index of alternate fuels ($PALT$).

Sales-for-resale demand equations are presented for each market region below. In the South Central, Southeast, and West regions, residential/commercial and industrial sales are aggregated together. Prices have been very low and demand has fluctuated considerably in these regions. Hence it is difficult to obtain stable estimates of demand elasticities on a disaggregated basis. In the West region, the price terms are statistically insignificant, but these variables with t -statistics slightly greater than one are retained in the equations (to show whatever price elasticity there is, whenever possible). As a general rule, variables with t -statistics greater than two are considered significant. All equations are estimated over the time bounds 1964 to 1970.

Northeast—residential/commercial:²⁶

$$\delta QSR_{CR} = -1587.65 - 9.117 \times 10^2 PSR + 5.371 \times 10^4 POIL + 63.33 \delta N \quad (31)$$

(0.04) (-1.21) (2.53) (5.33)

$$R^2 = 0.356 \quad F = 9.56 \quad S.E. = 2.47 \times 10^4$$

$$\text{Mean}(\delta QSR_{CR}) = 2.76 \times 10^4.$$

²⁶ Durbin-Watson statistics are not shown for the wholesale demand equations since data formatting problems made their calculation impossible here.

Northeast—industrial:

$$\begin{aligned}\delta QSRI &= 2.589 \times 10^4 - 1.079 \times 10^3 PSR \\ &\quad (0.81) \quad (-1.90) \\ &\quad + 3.057 \times 10^4 POIL + 1.975 VAM \\ &\quad (2.01) \quad (5.55) \quad (32) \\ R^2 &= 0.395 \quad F = 11.3 \quad S.E. = 1.87 \times 10^4 \\ \text{Mean}(\delta QSRI) &= 1.91 \times 10^4.\end{aligned}$$

North Central—residential/commercial:

$$\begin{aligned}\delta QSRCR &= 3.47 \times 10^4 - 1.108 \times 10^3 PSR + 14.43 \delta Y \\ &\quad (1.76) \quad (-2.10) \quad (11.56) \quad (33) \\ R^2 &= 0.725 \quad F = 78.9 \quad S.E. = 1.60 \times 10^4 \\ \text{Mean}(\delta QSRCR) &= 2.6 \times 10^4.\end{aligned}$$

North Central—industrial:

$$\begin{aligned}\delta QSRI &= 1.174 \times 10^4 - 9.65 \times 10^2 PSR \\ &\quad (0.51) \quad (-1.28) \\ &\quad + 8.192 \times 10^4 POIL + 39.25 K \\ &\quad (1.37) \quad (7.76) \quad (34) \\ R^2 &= 0.529 \quad F = 22.1 \quad S.E. = 1.90 \times 10^4 \\ \text{Mean}(\delta QSRI) &= 2.3 \times 10^4.\end{aligned}$$

West—total:

$$\begin{aligned}\delta QSRT &= -1860.89 - 2.287 \times 10^2 (PSR_{t-1} + PSR_{t-2})/2 \\ &\quad (0.23) \quad (-1.00) \\ &\quad + 9147.21 (POIL_{t-1} + POIL_{t-2})/2 + 26.70 VVC_{t-1} \\ &\quad (1.02) \quad (23.62) \quad (35) \\ R^2 &= 0.901 \quad F = 200.1 \quad S.E. = 1.67 \times 10^4 \\ \text{Mean}(\delta QSRT) &= 2.3 \times 10^4.\end{aligned}$$

South Central—total:

$$\begin{aligned}\delta QSRT &= 1.287 \times 10^6 - 1.389 \times 10^4 DDA \\ &\quad (1.40) \quad (-0.21) \\ &\quad - 5.332 \times 10^3 PSR + 276.67 K \\ &\quad (-1.50) \quad (5.20) \quad (36) \\ R^2 &= 0.803 \quad F = 51.8 \quad S.E. = 6.55 \times 10^4 \\ \text{Mean}(\delta QSRT) &= 1.008 \times 10^5.\end{aligned}$$

Southeast—residential/commercial:

$$\begin{aligned}\delta QSRCR &= 2.16 \times 10^4 - 1.612 \times 10^4 DDA - 5.841 \times 10^2 PSR \\ &\quad (1.93) \quad (-3.55) \quad (-2.31) \\ &\quad + 1.124 \times 10^4 POIL + 1.521 \delta Y \\ &\quad (2.21) \quad (1.26) \quad (37) \\ R^2 &= 0.267 \quad F = 4.1 \quad S.E. = 5.75 \times 10^3 \\ \text{Mean}(\delta QSRCR) &= 7.61 \times 10^3.\end{aligned}$$

Southeast—industrial:

$$\begin{aligned} \delta QSRI = & 2.260 \times 10^4 - 2.147 \times 10^4 DDA - 8.714 \times 10^3 PSR \\ & (1.35) \quad (-3.30) \quad (-2.01) \\ & + 6.012 \times 10^4 PALT + 2.086 VAM \\ & (2.85) \quad (1.67) \end{aligned} \quad (38)$$

$$R^2 = 0.373 \quad F = 6.55 \quad S.E. = 7.1 \times 10^3$$

$$\text{Mean}(\delta QSRI) = 1.34 \times 10^4.$$

The mainline industrial demand equation is shown below.

Mainline sales:

$$\begin{aligned} \delta QM = & 3.389 \times 10^4 - 1.079 \times 10^3 (PM_t + PM_{t-1})/2 \\ & (3.52) \quad (-3.59) \\ & + 3.447 \times 10^4 (PALT_t + PALT_{t-1})/2 \\ & (2.71) \end{aligned} \quad (39)$$

$$R^2 = 0.149 \quad F = 8.77 \quad S.E. = 1.79 \times 10^4$$

$$\text{Mean}(\delta QM) = 1.2 \times 10^4.$$

Finally, an intrastate demand equation has been estimated as follows:

$$\begin{aligned} \delta QINTRA = & 1.317 \times 10^4 - 10.85 (PM_t + PM_{t-1})/2 \\ & (2.19) \quad (-1.95) \\ & + 2.906 \delta Y \\ & (3.49) \end{aligned} \quad (40)$$

$$R^2 = 0.213 \quad F = 7.17 \quad S.E. = 2.02 \times 10^4$$

$$\text{Mean}(\delta QINTRA) = 7.23 \times 10^4.$$

These equations show negative price effects on demands, with statistically significant coefficients in six out of ten equations. The strongest effects were in the regions closer to producing centers (with more alternative sources of energy). Alternative fuel prices are important in determining mainline industrial sales. Size-of-market variables, such as consumer incomes or industrial investment or output, do not appear to be consistent causal factors in all sectors of the national gas market.

Interregional flows of gas

In order to calculate estimates of excess demand by consuming region (and underproduction by production district), we must have some estimate of how gas flows from the wellhead to the point of final consumption. We make estimates of interchange at an aggregate level using the demand regions, *W*, *NE*, *NC*, *SE*, *SC* and production regions, *P1*, *P2*, *P3*, *P4*, *P5*, *P6*, *P7*, *P8*, as defined below:

W—West: Arizona, California, Oregon, Washington, Idaho, Wyoming, Nevada, Utah, Colorado, New Mexico

- NE*—Northeast: New England, New York, New Jersey, Pennsylvania, Ohio, West Virginia, Kentucky, Maryland (and Delaware), Virginia
- NC*—North Central: South Dakota, Nebraska, Minnesota, Iowa, Missouri, Wisconsin, Illinois, Michigan, Indiana
- SE*—Southeast: Tennessee, Alabama, North Carolina, South Carolina, Georgia, Florida
- SC*—South Central: Kansas, Texas, Arkansas, Oklahoma, Louisiana, Mississippi
- P1*—Texas Railroad District 10, Oklahoma, Kansas
- P2*—New Mexico San Juan, New Mexico Permian, Texas 8, Texas 8A, Texas 7C
- P3*—Texas 1, Texas 9
- P4*—Texas 2, Texas 3, Texas 4, Texas 6, Louisiana North, Louisiana South, Mississippi
- P5*—Colorado, Wyoming
- P6*—Canada (exogenous production)
- P7*—California
- P8*—Kentucky, West Virginia.

The flows of gas from production region to consuming region as estimated for 1970 are shown in Table 2. In each box is an integer

TABLE 2
GAS FLOWS — 1970

SALES f_{jk} g_{jk}	W	NE	NC	SE	SC	TOTAL PRODUCTION
P1	235 0.069 0.055	82 0.020 0.019	1765 0.413 0.418	0 0.000 0.000	2136 0.247 0.506	4218 0.192
P2	1675 0.497 0.512	0 0.000 0.000	646 0 151 0.197	0 0.000 0.000	955 0.110 0.292	3276 0.149
P3	0 0.000 0.000	14 0.003 0.043	60 0.014 0.184	14 0.009 0.043	237 0.027 0.729	325 0.014
P4	0 0.000 0.000	3630 0.888 0.296	1776 0.416 0 145	1530 0.990 0.125	5293 0.607 0.428	12,229 0.558
P5	478 0.141 0.959	0 0.000 0.000	20 0.004 0.400	0 0.000 0.000	0 0.000 0.000	498 0.022
P6	305 0.090 1.00	0 0.000 0.000	0 0.000 0.000	0 0.000 0.000	0 0.000 0.000	305 0.013
P7	676 0.200 1.00	0 0.000 0.000	0 0.000 0.000	0 0.000 0.000	0 0.000 0.000	676 0.030
P8	0 0.000 0.000	359 0.087 1.00	0 0.000 0.000	0 0.000 0.000	0 0.000 0.000	359 0.016
TOTAL DEMAND	3369 0.153	4085 0.186	4267 0.194	1544 0.070	8621 0.393	21,886 1.000

which represents the flow from the particular production region (row) to the consuming region (column). The decimal to the left of this number (f_{jk}) represents the fraction of the consuming region's demand which comes from that production region, while the decimal to the right (g_{jk}) represents the fraction of gas from that production district going to a particular consuming region. Then, letting Q_j

equal production in district j and letting D_k equal demand in region k , we can calculate:

(1) $Q_j - \sum_k [f_{jk} D_k] =$ production shortage in producing region j
and

(2) $D_k - \sum_j [g_{jk} Q_j] =$ excess demand in consumption region k .

The numbers in Table 2 were derived from FPC data as follows:

- (1) The *FPC Sales by Producers of Natural Gas to Interstate Pipeline Companies* (1970) was used to determine for each pipeline company the distribution of purchases from each production region.
- (2) These purchases were then distributed to the various consuming regions (if they went to more than one) by using an estimation process based on the *FPC Form 2 Reports on Sales for Resale and Mainline Industrial Sales* by interstate pipeline companies and by using the FPC map, *Principal Natural Gas Pipelines in the United States 1968*. In many instances the companies' purchases and sales were divided into several nonoverlapping flows. These were then easy to determine. In others, it was necessary to estimate the actual flows, since more than one production region could be tapped for sales to more than one consuming region along the same pipeline. In these cases, judicious use of the Form 2 data was necessary to resolve the overlapping. It is believed that the resulting matrix is accurate enough to make estimates of excess demand.
- (3) The third step was to determine the size of intrastate and field sales²⁷ to each region. This was done by determining the difference between the total interstate sales coming from each production region and the value given for total production from that region by the AGA's *Reserves of Natural Gas*. The difference was then attributed to intrastate and field sales in the demand region containing the production region. In only one case (Permian Basin) was it necessary to split the intrastate production into two demand regions. This was accomplished by using the Bureau of Mines' consumption of natural gas data (Bureau of Mines, *Minerals Yearbook*) to determine total consumption for New Mexico. The quantity of intrastate sales necessary to bring New Mexico's interstate consumption up to the level of the Bureau of Mines' total consumption was added to the matrix. The remaining intrastate sales from the Permian Basin were then added to the South Central region.

This completed the flow table from which the fractions were easily calculated.

□ **Historic time period simulation of the model.** The model is used to predict the response over time of production supply and demand to a changing regulated field price. In order to help interpret (and also evaluate) these predictions, it is useful to examine a simulation of the model performed over an historic time period. If, for example, this

²⁷ Field sales are sales to an oil producer for pressuring purposes. They also include pipeline fuel.

simulation showed an increasingly upward bias in production over time, this could be taken into account when in model's predictions for future excess demand.

The model was simulated over the period 1966 to 1970, using historical values for all of the exogenous variables, including wellhead prices. The results of this simulation are shown in Table 3, which shows production supply and demand and for the wholesale price.

TABLE 3

HISTORIC SIMULATION OF PRODUCTION SUPPLY AND DEMAND

YEAR	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	SIMULATED PRODUCTION SUPPLY (TRILLIONS OF CU. FT.)	SIMULATED PRODUCTION DEMAND (TRILLIONS OF CU.FT.)	ACTUAL PRODUCTION (TRILLIONS OF CU. FT.)	SUPPLY ERROR [(1)-(3)]	DEMAND ERROR [(2)-(3)]	NEW CONTRACT FIELD PRICE (CENTS)	SIMULATED AVERAGE WHOLESALE PRICE (CENTS)	ACTUAL AVERAGE WHOLESALE PRICE (CENTS)
1966	19.4	15.6	17.5	1.9	-1.9	17.7	30.0	26.0
1967	19.6	16.3	18.4	1.2	-2.1	18.1	30.3	26.1
1968	19.8	17.1	19.4	0.4	-2.3	18.5	30.7	26.4
1969	20.0	17.9	20.7	-0.7	-2.8	19.3	31.2	27.5
1970	20.5	18.8	22.0	-1.5	-3.2	21.8	32.0	29.7
MEAN SUPPLY ERROR		= 0.3	MEAN DEMAND ERROR		= -2.5	MEAN WHOLESALE PRICE ERROR		
RMS SUPPLY ERROR		= 1.3	RMS DEMAND ERROR		= 2.5	RMS WHOLESALE PRICE ERROR		
MEAN ACTUAL PRODUCTION		= 19.6	MEAN ACTUAL WHOLESALE PRICE					

In addition to listing the simulated values, actual values, and errors for each variable, we calculated the mean and root-mean-square simulation errors.

Note that simulated production supply is larger than actual production levels in 1966, 1967, and 1968, but smaller than actual levels in 1969 and 1970, with a deficit of over two trillion cubic feet in 1970. Most of this error occurs in Louisiana and the Gulf of Mexico. Louisiana in the Outer Continental Shelf. There is more production forecast, probably because of a stepped-up program for Offshore Leasing. With a continuation of the usual Leasing forecasts in the 1970s should be more accurate. Production is underestimated slightly at first, but strongly later, that grows from about 1.9 trillion cubic feet in 1966 to 1.5 trillion cubic feet in 1969. At least part of the reason for this is the overestimate of the average wellhead price. The simulated wholesale price is about 4.2 cents too high in 1966 to 1968, and three cents too high in 1969 to 1970. This is due to more gas flowing from small production regions in the Gulf than in the real world (the small regions have higher mean wellhead prices), an error, while part of the *backcast*, would not appear in the *forecast* periods, because of depletion of reserves in the Gulf.

It would appear that the model's forecasts should have a margin of error (probably negative) of at least one trillion cubic feet. This should be taken into account in making the policy analyses that are presented in the next section. Excess demand may turn out to be somewhat more than what is forecast, and the price increases necessary to eliminate excess demand may actually be a few cents more than those which we pre-

■ With long lags from price changes to exploration, development, and production changes, it might be expected that very little additional supply would be forthcoming in the next five years from increased field prices. Given “roll-in” pricing in wholesale markets, field price increases would be passed through very gradually as wholesale price increases. Consequently, there might be only limited reductions in demand as a result of the price increases, as well. These possibilities can be investigated by completing several simulations with the model using alternative policies (in so far as these policies are reflected in alternative field prices), where the model has been set up to approximate roughly the conditions likely to be prevalent in the late 1970s.

The “gas economy” of the last half of the 1970s can be described in any number of ways, varying from quite expansive (because of shifts of energy demands to natural gas) to quite restrictive (because of reduced prices for alternative fuels as new coal-using technology is developed and as vastly increased supplies of oil resources become available). The expansive view, when introduced into the econometric model, would result in large increases of demands, while the more restrictive expectations would have quite the opposite effect. When these conditions are replicated in the model, by introduction of appropriate values of exogenous variables, the size of the shortage becomes either larger or smaller. The approach taken here is not to choose an extreme set of values of the exogenous variables to be inserted into the model; rather, we choose a set of values that follows from “median” conditions likely to prevail in energy markets in the near future.

The important exogenous determinants of demands for gas include state-by-state value added in manufacturing, population level, income, and capital equipment additions. It is forecast that value added, income per capita, and capital additions will grow at 4.2 percent per annum (based on the *Data Resources Quarterly Econometric Model* forecast for the period 1972 to 1980). These rates of growth are in terms of constant dollars; we have chosen a “median” expected rate of growth of prices close to 3.5 percent per annum (based on both the *Thurow-Ripley Long-Term Econometric Model* of Data Resources and the *DRI Quarterly Econometric Model*). Thus, value added and capacity grow at 7.7 percent in current dollar terms. It is assumed that the rate of growth of population will be limited to 1.1 percent per annum for the rest of the decade (in keeping with the assumptions used in the economy-wide models for generating the rates of growth of value added and capacity).

Other exogenous variables that determine the percentage share of total energy markets that will be realized by natural gas are: (1) wholesale prices of alternative fuels in residential and commercial consumption, and (2) wholesale prices of alternative fuels for industrial use. Both are expected to increase at 5.0 percent per annum in real terms and 8.5 percent per annum in current dollar terms. These exceed the forecasts of the National Petroleum Council,²⁸ but are close to an extrapolation of 1970 to 1973 price behavior that would essentially double oil prices over the 1972 to 1980 period.

Exogenous determinants of supply of reserves and production

²⁸ In [17].

include the field price for crude oil in the producing regions, and the average drilling costs for oil and gas together. It is expected that oil prices in the field will grow at the 8.5 percent rate per annum in current dollar terms (again exceeding the National Petroleum Council's predictions).²⁹ Average drilling costs are expected to increase at 6.8 percent per annum, in keeping with the trend of cost increases over the 1960s and early 1970s.

These values of the exogenous variables, together with assumed percentage increases of gas prices on new field contracts, can be inserted into the econometric model to produce simulated values of additions to reserves and to production by field region. Also, the model can produce simulations of demands for new production and for total production, and average wholesale prices for different classes of consumers in each of the wholesale markets. Each of the alternative regulatory policies implies a different series of field prices; separate simulations for each price series can be used to describe results of a policy and then can be compared with those for other policies.

□ **Partial deregulation of field prices.** A large number of alternative proposals have been made under the rubric of "deregulation of field prices." These have included complete and instantaneous deregulation on the one hand, and a slightly higher rate of increase of FPC-controlled prices on the other hand. But the central proposal is probably to allow new contract prices to seek their own levels, subject to a national ceiling on all new prices that would keep average wholesale prices from increasing by more than 50 percent over the subsequent five-year period. This policy has been simulated as a price increase of 15 cents per Mcf on all new contracts in 1974 over a 29.6 cents per Mcf price level on new contracts in 1973. Prices would increase an additional 3.0 cents per year after 1974, as part of the (allowed) response to rising drilling and operating costs.

The impact of the price increases on new discoveries of reserves would not occur in 1974, but rather would be spread over the period 1975 to 1978. It is expected that 1975 new discoveries would be slightly more than 9 trillion cubic feet—almost 30 percent more than the previous year (as shown in Table 4, for the econometric model simulation for "Deregulation"). By 1976 new discoveries would increase to the level of 11 trillion cubic feet, and by 1979 would be 19 trillion cubic feet. Total additions to reserves (including new discoveries, extensions, and revisions) would track new discoveries rather closely. A major part of the impact of price increases would be felt within four years, as an increase in annual additions 10 trillion cubic feet greater than previously experienced. The full impact would not be felt, however, until 1980, when total additions had leveled off at 34 trillion cubic feet, an amount approximately 21 trillion cubic feet greater than the 13 trillion realized before this series of price increases.

Production out of reserves can be expected to increase more quickly. The new contract field price increases would make it more profitable to increase the level of development of existing reserves. Over a longer period, the additions to total new reserves induced by price increases would provide more "reserve base" for more produc-

²⁹ *Ibid.*, pp. 40 *et seq.*

TABLE 4

ECONOMETRIC MODEL SIMULATION RESULTS FOR POLICY ALTERNATIVE "DEREGULATION"

YEAR	NEW DISCOVERIES (CONTINENTAL U.S., TRILLIONS OF CU.FT.)	TOTAL ADDITIONS TO RESERVES (CONTINENTAL U.S., TRILLIONS OF CU.FT.)	TOTAL RESERVES (CONTINENTAL U.S., TRILLIONS OF CU.FT.)	SUPPLY OF PRODUCTION (CONTINENTAL U.S., TRILLIONS OF CU.FT.)	DEMANDS FOR PRODUCTION (CONTINENTAL U.S., TRILLIONS OF CU.FT.)	EXCESS DEMAND FOR PRODUCTION (CONTINENTAL U.S., TRILLIONS OF CU.FT.)	NEW CONTRACT FIELD PRICE (CONTINENTAL U.S., CENTS PER Mcf)	AVERAGE WHOLESALE PRICE (CONTINENTAL U.S., CENTS PER Mcf)
1972	5.7	9.8	237.7	19.3	23.3	4.1	26.3	33.6
1973	6.1	12.7	229.0	22.1	24.4	2.3	29.6	36.7
1974	7.1	13.8	218.5	25.0	25.4	0.3	44.1	42.0
1975	9.1	15.4	208.8	26.0	26.4	0.3	47.7	45.8
1976	11.4	18.3	201.3	27.1	27.4	0.3	51.3	49.5
1977	14.1	22.2	196.5	28.2	28.5	0.3	54.9	53.3
1978	16.4	25.9	194.5	29.5	29.7	0.2	58.4	57.1
1979	19.1	29.9	195.3	31.0	31.0	0.0	62.0	61.0
1980	22.4	34.6	199.7	32.8	32.4	-0.3	65.5	64.9

tion. As a result, the supply of production would be expected to increase from 22 trillion cubic feet in 1973 to 32 trillion cubic feet by 1980. The increased production—mostly out of existing reserves—reduces the reserve-production ratio from 10:1 to slightly less than 7:1 at the lowest point. Price increases provide incentives to take out a larger proportion of reserves.

The demand side of gas markets experiences attrition as a result of the pass-through of the higher new contract field prices to the wholesale level. The wholesale prices do not increase very rapidly. Because of the large number of old contracts still outstanding at the end of the 1980s, the average wholesale price at that point in time falls short of the new contract field price level by 0.6 cents per Mcf. But wholesale prices rise on the average from the 30 cent level in the early 1970s to 64 cents by 1980, and the increases affect consumer demands. The demands for production by residential, commercial, and industrial users throughout the country rise from 24 trillion cubic feet in the early 1970s to only 32 trillion cubic feet by 1980. The expansive effects of increased investment, consumer incomes, and population are dampened by price increases. In the absence of the field and wholesale price increases, demands would have increased from 24 trillion cubic feet to 37 trillion cubic feet.³⁰ Therefore, as a result of an approximately 50 percent increase in wholesale prices spread over a 7-year period, *total* demands for production are decreased by 14 percent (from 37 to slightly more than 32 trillion cubic feet by 1980), and *additions* to demands are decreased by 40 percent (from 13 trillion to 8 trillion cubic feet).

The results of this policy of "partial deregulation" are to raise prices and to reduce greatly the magnitude of the gas shortage. Price increases, while substantial, eliminate the production shortage after five years, so that by 1979, the supply of production and the demand

³⁰ This estimate has been made by simulating with the same values of exogenous variables but with one cent annual field price increases after 1974, as in Table 5.

for production are both approximately 31 trillion cubic feet. In subsequent years, supplies again match demands rather closely; in fact, the small excess supplies (or negative deficits) would probably lead to rebuilding of the reserves base and to some price softening on new field contracts.

It might be asked whether the market clearing in these simulation results is highly sensitive to the choice of particular values of the exogenous variables. This would not seem to be the case. Simulations were also carried out with "high" and "low" sets of values of the exogenous variables; the "high" values included real rates of growth of value added, income, and plant and equipment outlay of 4.4 percent per annum, of population of 1.3 percent per annum, and real price increases for alternative fuels of 6.0 percent (residential and commercial use) and 6.0 percent (industrial use). These additions are matched by oil field price increases of 6 percent per annum in real terms. The rate of inflation was maintained at 3.5 percent per year. The result would be expected to increase demands for gas substantially—but also to increase supplies of both oil and gas (particularly associated gas reserves). The simulation with "high" values of the exogenous variables shows, in fact, production out of reserves to exceed 33.3 trillion cubic feet by 1980, because of the substantial additional volumes of new discoveries. Demands for reserves in this "high" value case exceed 33.7 trillion cubic feet so that markets come very close to clearing with production supplies and demands roughly equal in the last year of the decade.

"Low" values of the exogenous variables would be quite similar to the "median" values, except that value added and per capita incomes would increase one percent less per annum, and both field and wholesale prices of alternative fuels increase at 3.0 percent.⁸¹ In this simulation, the low values of field prices for oil dampen incentives for exploration of (associated) gas, so that annual new reserves by 1980 are approximately 3.2 trillion cubic feet less than shown in Table 4. The supply of production out of reserves is also 1.1 trillion cubic feet less, but demands for production are dampened from 32 trillion cubic feet in 1980 to 29 trillion cubic feet, so that more than enough production is supplied to meet demands over the period 1976 to 1980. Altogether, the shortage is eliminated in 1980 in the "high" value case and in 1975 in the "low" value case.

The impression is that "deregulation" that results in field prices in the 50 to 65 cent range over the rest of the decade would clear production markets of excess demand. There is some chance that there would still be a shortage, if "high" values of exogenous variables increased demands for gas more than reserves and production. But the most likely values of exogenous variables, even the "low" values, would clear production markets, with low reserve backing the only indicator of a gas shortage.⁸²

□ **Strict "cost of service" regulation of field prices.** One criticism of regulation is that prices have been allowed to creep up at the rate of

⁸¹ It is assumed that prices would increase at 3.5 percent per annum in money terms, in the "low" value case as well as the "median" and "high" value cases for simulation.

⁸² It might be possible to say that there is still a shortage. This is because the reserve-production ratio is as low as 7:1 when production supply equals produc-

2.5 cents per annum on new contracts, when there has been little “cost justification” for this to occur. It is argued that the Federal Power Commission should follow traditional procedures for setting prices—that is, full reviews of the historical costs of providing gas should be undertaken and prices should be limited to these costs. It is expected that under this regimen, price increases under new contracts would be limited to one cent per annum. This would result in average wholesale prices limited to the 40 to 48 cent range over the rest of the 1970s (as shown in Table 5).

TABLE 5
ECONOMETRIC MODEL SIMULATION FOR POLICY ALTERNATIVE “COST OF SERVICE REGULATION”

YEAR	NEW DISCOVERIES (CONTINENTAL U.S., TRILLIONS OF CU.FT.)	TOTAL ADDITIONS TO RESERVES (CONTINENTAL U.S., TRILLIONS OF CU. FT.)	TOTAL RESERVES (CONTINENTAL U.S., TRILLIONS OF CU. FT.)	SUPPLY OF PRODUCTION (CONTINENTAL U.S., TRILLIONS OF CU. FT.)	DEMANDS FOR PRODUCTION (CONTINENTAL U.S., TRILLIONS OF CU.FT.)	EXCESS DEMAND FOR PRODUCTION (CONTINENTAL U.S., TRILLIONS OF CU.FT.)	NEW CONTRACT FIELD PRICE (CONTINENTAL U.S., CENTS PER Mcf)	AVERAGE WHOLESALE PRICE (CONTINENTAL U.S., CENTS PER Mcf)
1972	5.7	9.8	237.7	19.3	23.3	4.0	26.3	33.5
1973	6.0	12.7	229.0	22.0	24.3	2.3	29.6	36.6
1974	7.1	13.8	220.8	22.8	25.6	2.8	30.5	38.4
1975	8.1	15.2	213.6	23.4	26.9	3.6	31.3	40.1
1976	9.2	16.8	207.5	24.0	28.5	4.5	32.1	41.7
1977	10.4	18.7	202.7	24.7	30.3	5.5	33.0	43.3
1978	11.9	20.8	199.3	25.7	32.2	6.6	33.8	44.8
1979	13.6	23.2	197.6	26.7	34.4	7.7	34.6	46.4
1980	15.7	26.3	197.5	28.0	36.9	8.9	35.5	48.0

The effects of such limited price increases on reserve and production supply would be substantial. New discoveries would increase by little more than one trillion cubic feet per year for the remainder of this decade, and total additions to reserves would be limited in 1980 to 26 trillion cubic feet per year. The supply of production out of reserves would range from 23 trillion cubic feet in 1974 to close to 28 trillion cubic feet in 1980. Altogether, additions to reserves are approximately 3.7 trillion cubic feet less per year than under the “deregulation” policy shown in Table 4.

The demands for production would be much greater when prices are held within this range. Total demands for new production increase by 1.3 to 2.5 trillion cubic feet per annum (as shown in Table 5). Total new demands by 1980 are expected to be more than 37 trillion cubic feet, 5 trillion cubic feet greater than under price increases associated with “deregulation.”

Excess demands persist over the remainder of the decade. Indeed, the gap between supply and demand for production increases as time passes, from approximately 3 trillion cubic feet in 1974 to 9 trillion cubic feet by 1980. The impact of this “cost of service regulation” policy is to exacerbate the excess demands that already exist,

tion demand—the backing for production is still very slim, imposing uncertainty on the buyers as to whether future production is forthcoming after gas-burning equipment has been installed. This is to say that, at equilibrium levels of production, there is still excess demand for reserves.

so that the shortage would be close to 25 percent of the total demands for production by 1980.

These conditions can be expected whether the "median" values, the "high" values, or the "low" values of the exogenous variables are used in the simulations. When "high" values are used, the size of excess demand in 1980 exceeds 9.6 trillion cubic feet per year, because additions to demands from market growth and price incentives exceed additions to (associated) supply from oil price increases. When "low" values are used, the excess demand is limited to 6.5 trillion cubic feet per year. But, in all three cases, the shortage increases over the remainder of the decade, and is equal to 20 percent of the demands for production.

□ **Maintenance of the regulatory "status quo."** Continuation of present regulation is envisioned as maintenance of the 1970 to 1972 policy of allowing price increases on new contracts of 2 to 4 cents per Mcf each year. The cases on field pricing brought before the Commission for the rest of the decade would be decided on any number of "cost" bases, as long as the result would be a price series for new commitments of reserves as shown in Table 6—with annual

TABLE 6

ECONOMETRIC MODEL SIMULATION RESULTS FOR POLICY ALTERNATIVE "REGULATORY STATUS QUO"

YEAR	NEW DISCOVERIES (CONTINENTAL U.S., TRILLIONS OF CU.FT.)	TOTAL ADDITIONS TO RESERVES (CONTINENTAL U.S., TRILLIONS OF CU.FT.)	TOTAL RESERVES (CONTINENTAL U.S., TRILLIONS OF CU.FT.)	SUPPLY OF PRODUCTION (CONTINENTAL U.S., OF CU.FT.)	DEMANDS FOR PRODUCTION (CONTINENTAL U.S., TRILLIONS OF CU.FT.)	EXCESS DEMAND FOR PRODUCTION (CONTINENTAL U.S., TRILLIONS OF CU.FT.)	NEW CONTRACT FIELD PRICE (CONTINENTAL U.S., CENTS PER Mcf)	AVERAGE WHOLESALE PRICE (CONTINENTAL U.S., CENTS PER Mcf)
1972	5.7	9.8	237.7	19.2	23.3	4.0	26.3	33.6
1973	6.0	12.7	228.9	22.1	24.3	2.3	29.6	36.7
1974	7.0	13.8	220.5	23.0	25.6	2.5	32.4	38.9
1975	8.3	15.2	212.8	23.9	26.9	2.9	35.1	41.2
1976	9.7	17.1	206.1	24.9	28.3	3.4	37.8	43.7
1977	11.3	19.4	200.8	26.0	29.9	3.9	40.6	46.2
1978	13.2	22.1	197.3	27.2	31.6	4.4	43.3	48.9
1979	15.5	25.4	195.8	28.6	33.4	4.8	45.9	51.7
1980	18.3	29.2	196.8	30.2	35.5	5.3	48.5	54.6

increments of approximately 3 cents per Mcf per year leading to new contract prices in 1980 of more than 50 cents per Mcf. Such price changes would not be easy to put into effect—each requiring multiple applications for certification of new contracts under conditions different from the previous year's certification and each requiring estimates of (increased) costs of providing the new contract production. The allowed increases—if the best that can be provided by a commission trying to reduce the shortage, but operating within orthodox regulatory "cost-pricing" procedures—would raise average wholesale prices from 37 cents to 55 cents by 1980. They would have a perceptible effect on both supplies of reserves and production, and demands for production.

Supplies of reserves are expected to increase at a slower rate under

these conditions than under "deregulation." The increments of new discoveries and total additions to reserves are shown in Table 6; total additions each year fall short of those for Table 4 by 5 to 6 trillion cubic feet at the end of the decade. The supply of production out of reserves is almost 3 trillion cubic feet less each year as a consequence. The combination of successive 3-cent increments leading to a higher final price, and of near-certainty of the occurrence of these increments, leads to some additions to both reserves and production. But these are not enough to eliminate the shortage.

Demands for production on the other hand are greater than under "deregulation." The successive increments of 3-cent new contract field price increases pass through as average wholesale price increases only gradually. The new contract field price is approximately 16 cents per Mcf higher in 1980 than in 1974, and the average wholesale price is about 15 cents higher. As a result, demands for production increase from 25 trillion cubic feet in 1974 to 35 trillion cubic feet in 1980. These demands at the end of the decade are 8 trillion cubic feet per annum greater than under the higher-priced "deregulation."

As a result of lower additions to supply and higher additions to demand, the shortage increases. In the mid-1970s the total supply of production would be from 2.5 to 3.5 trillion cubic feet less than the demand. By 1979 or 1980, further net increments of demand increase the shortage each year until it levels off at 5.3 trillion cubic feet. This short fall is a lower percentage of demand than under strict "cost-of-service" regulation (and the increments of supply are larger under the "status quo" than under "cost of service" in 1974 to 1980).

The use of "high" or "low" values of exogenous variables again does not affect the outcome of this policy. Markets are not cleared of the shortage under the "status quo" by using "low" values of exogenous variables, nor are they cleared by using "high" values of exogenous variables. The increment by which demands exceed the supply of production in 1980 equals 3.0 trillion cubic feet in the "low" simulation and 6.1 trillion cubic feet in the "high" simulation. In all of the simulations of continued regulation, policies of the Federal Power Commission that allow maximum legal price increases each year do not of themselves eliminate the shortage of production.

□ **Comparisons of alternative policies.** In comparing the alternative policies, it should be remembered that the econometric model presented in this paper is, of course, subject to some error, and as seen in the historic simulations, this error would most likely produce an underestimate of actual demands. Thus, the price increases consistent with our scenarios might actually be a few cents lower than those which we present here. Nonetheless, the qualitative results supporting deregulation would still hold. The government could "deregulate" by legislative act and reduce the production shortage to negligible size within four to five years. The FPC could not likely achieve the same result by stretching the standards for justifying price increases so as to allow increases of 3 cents per annum.

Both of the alternatives are preferable to strict "cost of service" regulation. Regulatory stringency is expected to increase the size of the shortage from approximately 3 trillion cubic feet of production per year at the present time to 9 trillion cubic feet per year by the

end of the decade (as in Table 5). Synthetic or liquefied gas plants which can produce at costs more than 65 cents per Mcf would be justifiable as investments under regulation, because they would ameliorate the shortage. That is, the gas industry would likely have to turn to more costly alternatives for synthetic supplies, *given the size of the "strict regulation" shortage*. Given strong political and economic reasons for reducing the shortage, the "strict regulation" way of doing so is going to be the most prolonged and the most costly.

Appendix

List of variables and data sources

■ The variables used to estimate the model, including sources of data, are listed below. The list is divided into field market variables and wholesale market variables. Following AGA procedures, all offshore data are included in the appropriate onshore district.

□ Wells.

- WXT*: Total exploratory wells, successful or otherwise.
- CWXT*: Cumulative value of exploratory wells (*WXT*), from 1963 to *t*.
- REVD*: Index of average deflated revenue from gas and oil for each production district.
- ATCD*: Average drilling cost per well of exploratory and development wells, from *Joint Association Survey*, (AGA/API/CPA).
- RISKV*: Index of the variance in size of discoveries over time by production district. This is a pure cross-section variable which does not change in time.
- AFX*: Average depth of an exploratory well (averaged over total wells).

- Reserves. All data from AGA/API/CPA *Reserves of Crude Oil*. Only available in disaggregated (i.e., associated-dissolved, nonassociated, by production district) form from 1966 except for year-end reserves (*YN*, *YA*, *US*) which we have from 1965. (Disaggregated data extrapolated for 1964 and 1965 to provide starting values for lagged and unlagged endogenous variables in the historical simulation.) Data are given in millions of cubic feet.

- SIZEDN*: Average size of nonassociated discoveries, by production district (averaged over total wells, dry and successful).
- SIZEDA*: Average size of associated discoveries, by production district (averaged over total wells, dry and successful).
- DDA*: Dummy variable for Louisiana South.
- DDB*: Dummy variable for the Permian producing region.
- DDC*: Dummy variable for Kansas, Oklahoma, TRRC Districts 1, 2, 3, 4, and 10.

DDD: Dummy variable for Louisiana North, Mississippi, New Mexico North, Pennsylvania, West Virginia-Kentucky, Wyoming, TRRC6 and TRRC9.

YN: Year-end nonassociated reserves. Reserves as defined by the AGA.

YA: Year-end associated-dissolved reserves. See *YN*.

YT: Year-end total gas reserves. $YT = YN + YA$.

YO: Year-end oil reserves.

US: Year-end reserves in underground storage other than in their original locations.

XN: Extensions of nonassociated gas. Includes any newly proved reserves already established in pools and fields.

XA: Extensions of associated-dissolved gas. See *XN*.

RN: Revisions of nonassociated gas. Includes any proved *decreases* in size of proved reserves discovered by drilling of extension or development wells or changes (+ or -) resulting from better engineering estimates of economically recoverable reserves in established pools.

RA: Revisions of associated-dissolved gas. See *RN*.

FN: New field discoveries of nonassociated gas (discovered by new field wildcats).

FA: New field discoveries of associated-dissolved gas. See *FN*.

PN: New pool discoveries of nonassociated gas.

PA: New pool discoveries of associated-dissolved gas.

DN: = $FN + PN$. Total new discoveries, nonassociated.

DA: = $FA + PA$. Total new discoveries, associated-dissolved gas.

] **Production.** All data are from AGA/API/CPA, *Reserves of Natural Gas*, and are available in disaggregated form from 1965 for gas (and 1940 for oil). In 10^6 cubic feet and 10^3 barrels.

QN: Production of nonassociated gas, i.e., net production.

QA: Production of associated-dissolved gas, net production.

QO: Production of oil.

QSRI: Sales for resale which will end up as industrial sales are determined by multiplying total sales for resale (from FPC Form 2 Reports) by the fraction of total industrial natural gas consumption (from U.S. Bureau of Mines' *Minerals Yearbook*).

QSRCR: The difference between total sales for resale and industrial sales for resale. (See *QSRI* for sources.)

QINTRA: The difference between total consumption of natural gas by state and total mainline sales plus sales for resale plus field sales in that state. Figures on total consumption by state are taken from Bureau of Mines'

Minerals Yearbook, as are total field sales. Mainline sales and sales for resale figures are from FPC Form 2 Reports.

QM: Total mainline industrial sales volume by interstate pipeline companies by state and year (1955 to 1970) in Mcf, from FPC Form 2 Reports.

QSRT: Total sales for resale volume by class A and B interstate pipeline companies, by state and year (1956 to 1970) in Mcf, from FPC Form 2 Reports.

L: Losses and waste, by states (Texas, Louisiana, and New Mexico undivided). From API *Petroleum Yearbook*.

□ Prices.

PG: New contract price of interstate sales of gas at the wellhead, from Table F, FPC, *Sales of Natural Gas*, for 1952 to 1971, in ¢/Mcf. (Compiled by Foster Associates, Inc.)

PO: Wellhead price of oil in \$/barrel for 1954 to 1971, from Bureau of Mines' *Minerals Yearbook*.

POIL: Average price in dollars per Mcf-energy-equivalent of fuel oil paid by electric power companies by state, from Edison Electric Institute, *Statistical Annual of the Electric Utility Industry*.

PALT: Price of alternate fuels in dollars per Mcf-energy-equivalent. Weighted average (over kilowatt-hours generated) of prices of fuel oil and coal consumed by electric utility industry in generating electric power, by state and year from 1962 to 1969. (See *POIL* for source.)

PM: Price of mainline industrial sales made on a firm basis, by state and year from 1952 to 1970. Data taken from FPC Form 2 Reports, in \$/Mcf.

PSR: Price of sales for resale in \$/Mcf from 1956 to 1970, by state and year. Data taken from FPC Form 2 Reports.

PRATIO: = PG/PO .

M: The distance from the center of a producing region to the population center of the consumption region.

V: The sum of the squares of pipeline diameters entering the state, by state.

VAM: Value added in manufacturing in millions of dollars by state and year, from 1958 to 1969 (interpolated for 1968). Data taken from U. S. Department of Commerce, Bureau of the Census, *Annual Survey of Manufacturers*.

K: New capital expenditures in millions of dollars; see *VAM* for source.

VCC: Value of construction contracts by states in which work was done, from 1956 to 1970, in millions of dollars. Data taken from *Statistical Abstract of the*

U.S. and from F. W. Dodge Corporation, *Dodge Construction Contract Statistics Service*.

Y: Personal income by state, from 1956 to 1969, in millions of dollars. Data taken from U. S. Bureau of Economic Analysis, *Survey of Current Business*.

N: Population by state and year, from 1955 to 1970, in millions. Data taken from U. S. Department of Commerce, Bureau of the Census, *Current Population Reports*.

References

1. ALCHIAN, A. "Costs and Outputs," in *Allocation of Economic Resources*, M. Abramowitz, ed., Stanford: Stanford University Press, 1958.
2. BALESTRA, P. *The Demand for Natural Gas in the United States*. Amsterdam: North Holland Publishing Company, 1967.
3. ——— AND NERLOVE, M. "Pooling Cross-Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas." *Econometrica*, Vol. 34, No. 3 (July 1966), pp. 585-612.
4. BREYER, S. G. AND MACAVOY, P. W. *Energy Regulation by the Federal Power Commission*. Washington, D. C.: The Brookings Institution, forthcoming, Winter 1973.
5. ———. "The Natural Gas Shortage and the Regulation of Natural Gas Producers." *Harvard Law Review*, Vol. 86, No. 6 (April 1973), pp. 941-987.
6. CHETTY, V. K. "Pooling of Time Series and Cross-Section Data." *Econometrica*, Vol. 36, No. 2 (April 1968), pp. 279-290.
7. CHOW, G. C. "Technological Change and the Demand for Computers." *The American Economic Review*, Vol. 57, No. 5 (December 1967), pp. 1117-1130.
8. ERICKSON, E. W. AND SPANN, R. M. "Supply Price in a Regulated Industry, The Case of Natural Gas." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 94-121.
9. FISHER, F. M. *Supply and Costs in the U. S. Petroleum Industry: Two Econometric Studies*. Baltimore: Johns Hopkins Press, 1964.
10. GARFIELD, P. J. AND LOVEJOY, W. F. *Public Utility Economics*. Englewood, N. J.: Prentice-Hall, 1964.
11. GARLAND, T. M. AND DIETZMAN, W. D. "Engineering Cost Study of Development Wells and Profitability Analysis of Crude Oil Production." U. S. Bureau of Mines Information Circular #8561, 1972.
12. HOUTHAKKER, H. S. AND TAYLOR, L. D. *Consumer Demand in the United States 1929-1970*. Cambridge: Harvard Univ. Press, 1966.
13. KAUFMAN, G. "Sampling without Replacement and Proportional to Random Size." Memorandum II, to the M.I.T. Natural Gas Research Group, March 19, 1973.
14. KHAZZOOM, J. D. "The F.P.C. Staff's Econometric Model of Natural Gas Supply in the United States." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 51-93.
15. MACAVOY, P. W. *Price Formation in Natural Gas Fields*. New Haven: Yale Univ. Press, 1962.
16. ———. "The Regulation-Induced Shortage of Natural Gas." *Journal of Law and Economics*, Vol. 14, No. 1 (April 1971), pp. 167-199.
17. NATIONAL PETROLEUM COUNCIL. *United States Energy Outlook*. Washington, D. C.: December 1972.
18. RYAN, J. M. "National Academy of Sciences Report on Energy Resources: Discussion of Limitations of Logistic Projections." *Bulletin of the American Association of Petroleum Geologists*, Vol. 49, No. 10 (October 1965), pp. 1713-1727.
19. U. S. FEDERAL POWER COMMISSION. "Hugeton-Anadarko Area Rate Proceeding." *Annual Report of the Federal Power Commission*, 44 FPC 761, 1970.

20. ———. (Bureau of Natural Gas.) *Natural Gas Supply and Demand, 1971–1990*. Washington, D. C.: U. S. Government Printing Office, 1972.
21. ———. *Proceedings on Curtailment of Gas Deliveries of Interstate Pipelines*. Washington, D. C.: U. S. Government Printing Office, 1972.
22. ———. "Southern Louisiana Area Rate Proceeding." *Annual Report of the Federal Power Commission*, 46 FPC 86, 1971.
23. U. S. SENATE COMMERCE COMMITTEE. "Proposed Amendments to the Natural Gas Act." Staff Memorandum to the Chief Counsel, April 24, 1973.
24. WELLISZ, S. H. "Regulation of Natural Gas Pipeline Companies: An Economic Analysis." *Journal of Political Economy*, Vol. 71, No. 1 (February 1963), pp. 30–43.
25. WESTFIELD, F. M. "Methodology of Evaluating Economic Regulation." *The American Economic Review*, Vol. 61, No. 2 (May 1971), pp. 211–217.
26. WINCH, D. M. "Consumers' Surplus and the Compensation Principle." *The American Economic Review*, Vol. 55, No. 4 (September 1965), pp. 395–423.

Additive nonhomogeneous production functions in telecommunications

Ephraim F. Sudit

Corporate Planning Organization
American Telephone and Telegraph Company

An additive form of a nonhomogeneous production function is suggested in this paper as a tool to study key economic trends in production processes. This formulation allows for variable elasticities of substitution and for variable economies of scale, and is readily estimated by ordinary least squares. The marginal products are explicitly stated as a function of the input ratios. This additive nonhomogeneous functional form has been empirically tested on telecommunications data, related to Bell System, Western Electric, and Bell Canada production systems. Marginal products and elasticities have been computed and compared with the same series generated by a multiplicative nonhomogeneous form recently proposed by H. D. Vinod. Several similarities as well as major differences between the findings are cited and discussed. The author concludes that the additive formulation offers a more plausible description of some aspects of the Bell System and Bell Canada production processes than comparable findings generated by the multiplicative formulation.

■ The purpose of this paper is to offer an alternative nonhomogeneous formulation to H. D. Vinod's multiplicative nonhomogeneous production function.¹ This proposed additive nonhomogeneous production function seems to offer economic properties which are somewhat more consistent with fundamental postulates of production theory, as well as a better empirical description of several telecommunication production processes.

Section 2 discusses the mathematical properties of the additive nonhomogeneous function (*A-NH*) and compares them with the properties of Vinod's formulation. It is shown that, while the *A-NH* is as nonrestrictive in its underlying assumptions as Vinod's mul-

1. Introduction

Ephraim F. Sudit received his B.A. (1966) in Economics and Political Science from the Hebrew University, Jerusalem, and his M.B.A. (1967) from Columbia University. Presently he is a Ph.D. candidate at the Graduate School of Business, New York University. He currently holds the position of Supervisor—Productivity Studies in the Economic Analysis Section at AT&T and is engaged in productivity and production analysis.

This paper has benefited greatly from the work of H. D. Vinod of Bell Labs as well as from numerous extremely helpful suggestions by David K. Whitcomb of New York University and Charles Shami of Yeshiva University and AT&T. I was also helped by valuable comments on earlier versions by Malcolm Burnside and Kenneth Militzer of AT&T, Michael Keenan, and Samprit Chatterjee of New York University, the referee and the editors of the *Bell Journal*, as well as by the computer assistance of Vernelle Soper of AT&T.

The views expressed in this paper do not reflect in any way the views of AT&T nor of any Bell System employees but myself.

¹ See [6], pp. 531–543.

multiplicative nonhomogeneous function (*M-NH*), it explicitly sets the marginal products of the factors as functions of the ratio of the inputs. The *A-NH* also produces a more straightforward and interesting expression for scale elasticity. The elasticity of factor substitution formula in the *A-NH* case, however, does not reduce itself to the neat elasticity of substitution formula of the *M-NH*.

Section 3 exhibits the empirical results of fitting the *A-NH* formulation to aggregate Bell System and Bell Canada data, as well as to Western Electric's microeconomic data. The results are compared with Vinod's *M-NH* empirical findings for the same bodies of data.² Major differences in implied economic behavior are discovered and discussed in detail.

2. The additive nonhomogeneous formulation

■ Consider the additive nonhomogeneous formulation:

$$y = a_1x_1 + a_2x_2 + a_3x_1 \ln x_2 + a_4x_2 \ln x_1, \quad (1)$$

where Y is output, X_1 and X_2 are labor and capital inputs respectively, and the a s denote parameters.³ It follows that the marginal products of labor and capital, f_1 and f_2 , respectively, are:

$$f_1 = \frac{\partial y}{\partial x_1} = a_1 + a_3 \ln x_2 + a_4 \frac{x_2}{x_1} \quad (2)$$

$$f_2 = \frac{\partial y}{\partial x_2} = a_2 + a_3 \frac{x_1}{x_2} + a_4 \ln x_1. \quad (3)$$

Expressions (2) and (3) state explicitly that the marginal product of a factor is a function of the ratio of the quantities of the factors, as well as the logarithmic function of the other input. Thus, if a_3 and a_4 are positive, the marginal product of labor, for instance, is directly related to the capital intensity of the production process (capital-labor ratio), and logarithmically related to the level of the capital input. This of course amounts to a straightforward exposition of the rationale behind the postulate of diminishing returns.⁴

If $a_3 > 0$ and $a_4 > 0$, the *A-NH* function exhibits diminishing returns with respect to both inputs since

$$f_{11} = \frac{\partial^2 y}{\partial x_1^2} = -a_4 \frac{x_2}{x_1^2} \quad (4)$$

$$f_{22} = \frac{\partial^2 y}{\partial x_2^2} = -a_3 \frac{x_1}{x_2^2}. \quad (5)$$

It is clear, however, from (4) and (5) that the inputs of the *A-NH* function are not necessarily restricted to diminishing returns. One or both inputs could well exhibit increasing returns.

The shift in the marginal product of one input in response to a

² *Ibid.*

³ The function in (1) can be easily generalized to cases of more than two inputs.

⁴ By contrast, the *M-NH* formulation generates simple marginal elasticity formulae, but the marginal product expressions are relatively complex for a straightforward deduction of simple relationships. See [6], p. 533.

change in another input could be expressed as:

$$f_{12} = \frac{\partial^2 y}{\partial x_1 \partial x_2} = \frac{a_3}{x_2} + \frac{a_4}{x_1}. \quad (6)$$

According to Frisch,⁵ f_{12} is an indicator of whether the inputs are complementary or competitive with respect to each other. Expressions f_{11} and f_{22} are termed direct product accelerations of labor and capital, respectively; f_{12} is termed cross product acceleration.

The scale elasticity ϵ (the sum of the marginal elasticities of the products) can be derived as follows:

$$\begin{aligned} \epsilon &= \epsilon_1 + \epsilon_2 = \frac{1}{y} \left(x_1 \frac{\partial y}{\partial x_1} + x_2 \frac{\partial y}{\partial x_2} \right) \\ &= \frac{1}{y} (a_1 x_1 + a_3 x_1 \ln x_2 + a_2 x_2 + a_4 x_2 \ln x_1 + a_3 x_1 + a_4 x_2) \\ &= 1 + \frac{a_3 x_1 + a_4 x_2}{y}. \quad (7) \end{aligned}$$

The scale elasticity is defined as the percentage change in putput associated with a simultaneous infinitesimal percentage change equal to all inputs. According to (7), ϵ can assume any value, which means that the "localized" returns to scale are variable over the scale of production. The conditions, $a_3 \geq 0$ and $a_4 \geq 0$ assure $\epsilon \geq 1$, although these are only sufficient, but not necessary, conditions. Thus, if $a_3 = 0$ and $a_4 = 0$, then $\epsilon = 1$ (constant returns to scale), and the production function is reduced to the linearly homogeneous additive formulation $y = a_1 x_1 + a_2 x_2$.

The general condition for the convexity of the two factor isoquant with respect to the origin is known to be:

$$(f_2^2 f_{11} - 2 f_1 f_2 f_{12} + f_1^2 f_{22}) < 0. \quad (8)$$

The A - NH isoquants are not constrained to a convex relationship with respect to the origin. Their slopes can be easily derived from (1) as:

$$\frac{dx_1}{dx_2} = - \frac{\left(a_2 + a_3 \frac{x_1}{x_2} + a_4 \ln x_1 \right)}{\left(a_1 + a_4 \frac{x_2}{x_1} + a_3 \ln x_2 \right)}. \quad (9)$$

If both factors exhibit diminishing returns ($f_{11} < 0$ and $f_{22} < 0$), and both are substitutes and complements at the same time (i.e., $f_{11} > 0, f_{22} > 0$ and $f_{12} > 0$),⁶ then (8) will be negative and the isoquant will be convex to the origin. However, such a combination constitutes only a sufficient and not a necessary condition for convexity. Inputs constrained by the A - NH formulation will exhibit diminishing returns, if and only if $a_3 > 0$ and $a_4 > 0$. The latter assures a case where the A - NH isoquants tend to be asymptotic to the $X_1 = 1$ and

⁵ See [5], pp. 58-61.

⁶ Our concepts of substitutability and complementarity among factors of production are applied strictly in the technical sense, without resorting to a market environment, and correspond to Frisch's definitions in [5], pp. 55-61.

$X_2 = 1$ lines. This follows immediately from (9). If $a_3 > 0$ and $a_4 > 0$ then when $X_1 \rightarrow 1$ from above and $X_2 \rightarrow \infty$, the numerator of (9) approaches the value of a_2 on account of $\frac{x_1}{x_2}$ and $\ln x_1$'s approaching zero. The denominator becomes very large on account of $\frac{x_2}{x_1}$ and $\ln x_2$'s approaching infinity. Thus, $\frac{dx_1}{dx_2} \rightarrow 0$. By the same logic, when $X_2 \rightarrow 1$ from above and $X_1 \rightarrow \infty$, $\frac{dx_1}{dx_2} \rightarrow \infty$.

From the foregoing discussion, it is clear that the elasticity of factor substitution for the *A-NH* is not restricted to any constant value. The *A-NH* thus belongs to the variable elasticity of substitution class (*VES*) of production functions. It is noteworthy that in conventional formulations such as the Cobb-Douglas or the *CES*, where σ is defined independently of any particular factor combination, there is no unique way of defining the elasticity of substitution for formulations of the *A-NH* and *M-NH* type, since it is a function of scale and factor combinations.

□ **Estimation.** The *A-NH* formula in (1) is linear in its parameters, and therefore can be readily estimated by ordinary least squares. Using the estimates of the parameters, the series of the marginal products and marginal elasticities could be directly computed according to expressions (2) and (3). The elasticity of factor substitution, σ , could be numerically computed by using the well-known general expression:

$$\sigma_{12} = \frac{f_1 f_2 (x_1 f_1 + x_2 f_2)}{-x_1 x_2 (f_{11} f_2^2 - 2 f_{12} f_1 f_2 + f_{22} f_1^2)}. \quad (10)$$

3. Empirical findings

■ The additive nonhomogeneous production function formulation has been tested on various bodies of data with encouraging economic and statistical results. We shall report below on results generated by applying the *A-NH* to three sets of telecommunications production data: (1) Bell System aggregate production series for the postwar period; (2) Western Electric data on the manufacture of "Sealed Contact" units; and (3) Bell Canada aggregate production series for the 1952 to 1967 period. The two former bodies of data are the same as those used by Vinod⁷ to test his *M-NH* function. Thus, *A-NH* and *M-NH* empirical performance can be directly compared.

□ ***A-NH* production function for the Bell System.** Exploratory aggregate annual production times series for the Bell System for the 1947 to 1971 period were used for estimation. Let Y = Bell System net value added in 1967 constant dollars, X_1 = Bell System weighted man-hours, X_2 = Bell System net capital stock in 1967 constant dollars. The *A-NH* production function, with t -statistics in paren-

theses, is estimated as:

$$\begin{aligned}
 y = & 27.9844x_1 - 1.47495x_2 - 3.12076x_1 \ln x_2 + 0.26230x_2 \ln x_1 \\
 & (5.17597) \quad (-10.0271) \quad (-5.10746) \quad (12.0589) \quad (11) \\
 R^2 = & 0.9950 \quad S.E. = 201.97 \quad DW = 1.0998 .
 \end{aligned}$$

Prior to any detailed analysis of these findings, it should be noted that technological change variables were not explicitly and separately specified in model (11). Consequently, the impact of technological changes is likely to be embedded in the capital and labor parameters. Ideally, aggregate indexes of technological change, of both embodied and disembodied nature, should have been specified as inputs in model (11). However, available technological data in the present form did not lend themselves, in the opinion of the author, to a reasonably meaningful aggregate representation. Alternatively, conventional proxies divorced of a hard engineering data base were considered grossly inaccurate and potentially misleading. Therefore, at this phase of the research, a presentation of a model with an acknowledged possible misspecification was considered preferable to a presentation with a misleading appearance of correct specification.

TABLE 1
BELL SYSTEM MARGINAL PRODUCTS AND ELASTICITIES, 1947 TO 1971
(ADDITIVE NONHOMOGENEOUS PRODUCTION FUNCTION)

YEAR	MARGINAL PRODUCT OF LABOR	MARGINAL PRODUCT OF CAPITAL	MARGINAL ELASTICITY OF LABOR	MARGINAL ELASTICITY OF CAPITAL	SCALE ELASTICITY	ELASTICITY OF FACTOR SUBSTITUTION
	f_1	f_2	ϵ_1	ϵ_2	ϵ	σ
1947	\$ 2.912	\$ 0.007	1.093	0.037	1.130	-0.032
1948	2.822	0.043	1.039	0.229	1.268	-0.288
1949	2.820	0.064	1.036	0.371	1.407	-0.726
1950	2.882	0.077	0.970	0.432	1.402	-1.626
1951	2.717	0.086	0.855	0.445	1.301	-2.963
1952	2.623	0.108	0.809	0.571	1.380	2.936
1953	2.571	0.128	0.787	0.709	1.495	1.204
1954	2.638	0.140	0.788	0.804	1.592	0.908
1955	2.849	0.152	0.768	0.860	1.628	0.746
1956	2.588	0.173	0.693	0.970	1.662	0.612
1957	2.938	0.180	0.709	1.005	1.714	0.580
1958	3.175	0.191	0.739	1.107	1.846	0.546
1959	3.677	0.192	0.753	1.083	1.836	0.546
1960	3.901	0.198	0.741	1.088	1.829	0.535
1961	4.231	0.203	0.755	1.120	1.876	0.527
1962	4.526	0.210	0.761	1.162	1.922	0.517
1963	4.843	0.216	0.767	1.186	1.953	0.513
1964	4.925	0.226	0.742	1.220	1.961	0.450
1965	5.086	0.239	0.729	1.275	2.003	0.488
1966	4.987	0.250	0.661	1.238	1.898	0.476
1967	5.203	0.257	0.642	1.233	1.875	0.473
1968	5.280	0.263	0.601	1.189	1.790	0.469
1969	5.139	0.277	0.551	1.183	1.734	0.455
1970	5.141	0.291	0.539	1.238	1.777	0.445
1971	5.593	0.299	0.565	1.304	1.869	0.449

The numerical series of the marginal products and marginal elasticities of labor and capital as well as the scale elasticities and the elasticities of factor substitution are shown in Table 1. The product accelerations are shown in Table 2.

TABLE 2

BELL SYSTEM PRODUCT ACCELERATIONS, 1947 TO 1971
(ADDITIVE NONHOMOGENEOUS PRODUCTION FUNCTION)

YEAR	DIRECT ACCELERATION (LABOR)	DIRECT ACCELERATION (CAPITAL)	CROSS ACCELERATION (LABOR-CAPITAL)
	f_{11}	f_{22}	f_{12}
1947	\$ -0.484×10^{-2}	\$ 0.258×10^{-4}	\$ 0.331×10^{-4}
1948	-0.513×10^{-2}	0.201×10^{-4}	0.635×10^{-4}
1949	-0.543×10^{-2}	0.168×10^{-4}	0.833×10^{-4}
1950	-0.576×10^{-2}	0.148×10^{-4}	0.989×10^{-4}
1951	-0.541×10^{-2}	0.146×10^{-4}	0.906×10^{-4}
1952	-0.540×10^{-2}	0.127×10^{-4}	0.966×10^{-4}
1953	-0.545×10^{-2}	0.110×10^{-4}	0.103×10^{-3}
1954	-0.571×10^{-2}	0.967×10^{-5}	0.113×10^{-3}
1955	-0.629×10^{-2}	0.816×10^{-5}	0.130×10^{-3}
1956	-0.578×10^{-2}	0.755×10^{-5}	0.119×10^{-3}
1957	-0.656×10^{-2}	0.638×10^{-5}	0.138×10^{-3}
1958	-0.707×10^{-2}	0.544×10^{-5}	0.148×10^{-3}
1959	-0.814×10^{-2}	0.464×10^{-5}	0.168×10^{-3}
1960	-0.858×10^{-2}	0.419×10^{-5}	0.175×10^{-3}
1961	-0.922×10^{-2}	0.370×10^{-5}	0.183×10^{-3}
1962	-0.973×10^{-2}	0.326×10^{-5}	0.189×10^{-3}
1963	-0.103×10^{-1}	0.292×10^{-5}	0.194×10^{-3}
1964	-0.103×10^{-1}	0.267×10^{-5}	0.191×10^{-3}
1965	-0.103×10^{-1}	0.237×10^{-5}	0.189×10^{-3}
1966	-0.997×10^{-2}	0.226×10^{-5}	0.182×10^{-3}
1967	-0.102×10^{-1}	0.206×10^{-5}	0.182×10^{-3}
1968	-0.102×10^{-1}	0.195×10^{-5}	0.180×10^{-3}
1969	-0.971×10^{-2}	0.184×10^{-5}	0.171×10^{-3}
1970	-0.945×10^{-2}	0.169×10^{-5}	0.165×10^{-3}
1971	-0.993×10^{-2}	0.147×10^{-5}	0.168×10^{-3}

With respect to the estimated coefficients in (11), $a_8 = -3.12076 < 0$ immediately implies that the capital factor exhibits increasing returns to scale. However, these increasing returns to capital are almost negligible throughout the postwar period as shown in Table 2. In 1971 for instance, an increase of one 1967 dollar of Bell System capital stock induced only 0.00015¢ increase in the marginal product of capital. This minor acceleration may be explained on the grounds that our specification does not include a separate variable for technological change. Technological advances in the Bell System during the postwar period are believed to have been largely capital-embodied and labor-saving. The introduction of Direct Distance Dialing (DDD) is but one example. Thus, it may be expected that the increase in capital stock over time, coupled with far-reaching technological innovations made possible through the new capital equipment introduced, may generate, slightly increasing, rather than diminishing, returns to capital. On the other hand, it may be argued that since Bell System capital input is derived by repricing past vintages year by year—thereby reflecting shifting age distributions—embodied technological changes, insofar as they can be measured by changes over time in the vintage composition of the capital stock, have in fact been already taken into account by the variations in the capital input measure.

These phenomena are reflected in the trends in the marginal products and marginal elasticities of the inputs as shown in Table 1. The marginal product of labor almost doubled during the observed

period, from around \$2.8 of revenues in 1967 dollars per incremental man-hour in the late 1940s, to approximately \$5.6 of revenues in 1967 dollars per incremental man-hour in 1971. The marginal product of capital, however, increased at a much faster rate. In 1950, for example, an incremental 1967 dollar of net capital stock produced approximately 8¢ of additional 1967 dollar revenues. In 1971, an incremental 1967 dollar of net capital stock produced approximately 30¢ of additional 1967 dollar revenues. This divergence in the trend is even more apparent in comparing the relative trends in the marginal elasticities. In 1950 a one-percent increase in labor input generated a 0.97-percent increase in output, while a one-percent increase in the capital input induced only 0.43-percent increase in output. In 1967 the relative magnitudes of the elasticities were reversed: one-percent increase in the labor input generated only 0.56-percent increase in output, while one-percent increase in the capital input induced 1.3-percent increase in output.

The scale elasticity movements, also shown in Table 1, present an interesting picture. Bell System scale elasticity peaked in 1965 when a one-percent simultaneous increase in the inputs was associated with a 2-percent increase in output. During the 1966 to 1969 period Bell System scale elasticity declined from its 1965 peak, with an upward movement resuming in 1970 to 1971. On the whole however, the postwar Bell System data exhibit considerable economies of scale. Which part of these impressive scale economies was due solely to technological advance, it is impossible to tell. The elasticity of factor substitution was negative for the 1947 to 1951 years. Since 1952, the values of σ were positive and continuously declining, except for a minor reversal of the trend in 1971. For the period 1954 to 1971 the elasticity of substitution exhibited values lower than unity, and for the years 1957 to 1971, the values of σ were in the 0.6–0.4 range.

The levels of product acceleration (Table 2) show very small diminishing return impact on the labor side, negligible increasing returns to capital, and a minimal interdependence effect between the marginal product of one input and the level of the other. The latter result suggests that Bell System labor and capital are in effect neither complements nor competitors.

The considerable variability in the Bell System scale elasticity over the postwar years suggests that Bell System production processes diverge from a homogeneous pattern. Likewise, the variability in Bell System postwar elasticity of factor substitution indicates that its production process is not characterized by a constant elasticity of substitution.

□ **Bell System labor market.** The Bell System value of the marginal product of labor is obtained by multiplying the marginal product of labor series by an index of telephone rates and by comparing it to the Bell System cost per unit of labor.⁸ The value of the marginal product of labor is derived by multiplying the marginal product of labor series, generated by the *A-NH* function (Table 1), by the Bell System composite rate index. The Bell System unit labor cost is

TABLE 3
BELL SYSTEM VALUE OF THE
MARGINAL PRODUCT OF LABOR
VS. LABOR UNIT COST
(ADDITIVE NONHOMOGENEOUS
PRODUCTION FUNCTION)

YEAR	VALUE OF THE MARGINAL PRODUCT OF LABOR	TOTAL EMPLOYMENT COST PER MANHOUR
	VMP _L	W
1947	\$ 2.323	\$ 1.789
1948	2 329	1.952
1949	2.415	2.051
1950	2 561	2.130
1951	2.458	2.253
1952	2.451	2.389
1953	2.482	2.474
1954	2.630	2.561
1955	2.862	2.735
1956	2.614	2.730
1957	2.988	2 851
1958	3.286	2.853
1959	2.819	3.089
1960	4.034	3 223
1961	4.382	3.350
1962	4.677	3.459
1963	4 980	3.575
1964	5 055	3.760
1965	5.139	3.901
1966	5.015	4.089
1967	5.203	4.324
1968	5.250	4.526
1969	5.152	4 981
1970	5.220	5.440
1971	5 891	6.031

⁸ The well-known optimization condition in the labor market is $f_L P = W$, where P is a price of a unit of output, and W is the wage rate.

computed as the total employment cost per weighted man-hour. The resulting series are shown in the Table 3 and plotted on Figure 1.

FIGURE 1

BELL SYSTEM VALUE OF THE MARGINAL PRODUCT OF LABOR VS. LABOR UNIT COSTS

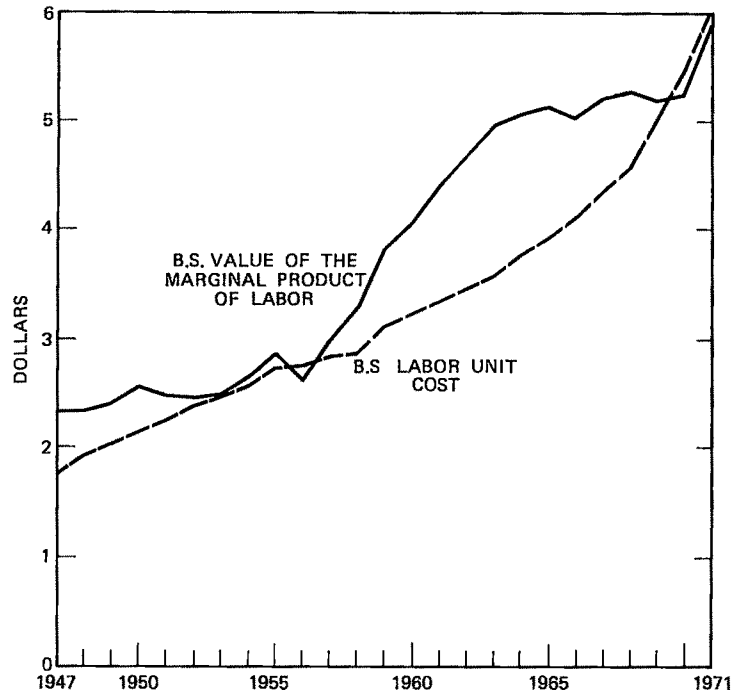


TABLE 4

BELL SYSTEM VALUE OF THE MARGINAL PRODUCT OF CAPITAL VS. CAPITAL UNIT COST (ADDITIVE NONHOMOGENEOUS PRODUCTION FUNCTION)

YEAR	VALUE OF THE MARGINAL PRODUCT OF CAPITAL	COST PER UNIT OF CAPITAL
	VMP _L	i
1947	\$ 0.006	\$ 0.054
1948	0.035	0.054
1949	0.055	0.055
1950	0.068	0.060
1951	0.078	0.060
1952	0.101	0.063
1953	0.124	0.068
1954	0.140	0.069
1955	0.153	0.075
1956	0.175	0.078
1957	0.184	0.078
1958	0.198	0.079
1959	0.200	0.083
1960	0.204	0.084
1961	0.210	0.085
1962	0.217	0.086
1963	0.222	0.087
1964	0.232	0.091
1965	0.241	0.095
1966	0.251	0.100
1967	0.257	0.101
1968	0.262	0.099
1969	0.278	0.102
1970	0.296	0.100
1971	0.315	0.092

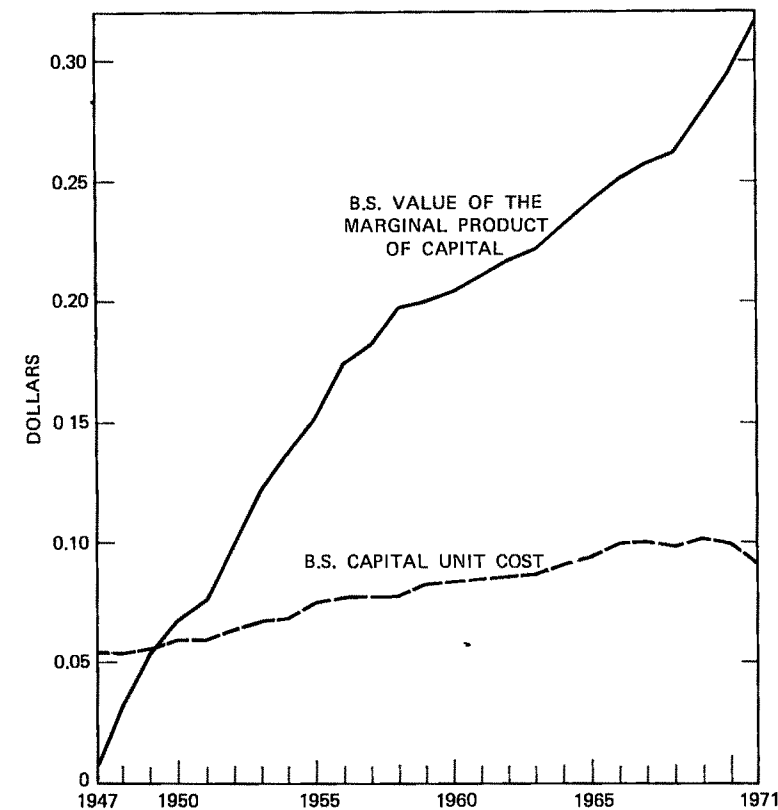
Considering the stochastic nature of the value of the marginal product of the labor series, the closeness of the two series is startling. This numerical correspondence strongly suggests that in most of the postwar years the Bell System was relatively close to the economically optimal pattern of hiring and paying labor according of the value of its marginal product.⁹ For instance, in 1971 Bell System compensation per man-hour (\$6.03) was practically equal to the value of the additional output of an incremental man-hour (\$5.89).

□ **Bell System capital market.** The Bell System value of the marginal product of capital can be compared with its cost per unit of capital. The value of the marginal product of capital is obtained by multiplying the marginal product of capital, as generated by the *A-NH* function in Table 1, by the index of telephone rates.¹⁰ The capital cost per unit is computed by weighting the opportunity cost of equity and the marginal cost of debt by the equity-debt shares in total obligations. The two series for the postwar period are shown in Table 4 and plotted in Figure 2. They indicate that, except for the late 1940s, the value of the marginal product of capital exceeded by far the capital cost per unit. The gap was widened consistently over the years. In

⁹ Although significant differences of the magnitude of \$1 per man-hour occurred in several years, and these might be viewed as lags, it must be remembered that the value of the marginal product of labor is not deterministic, and confidence limits ought to be considered before precise conclusions are drawn.

¹⁰ The well-known optimization condition in the capital market is $f_z P = i$, where i is the cost per unit of capital.

FIGURE 2
 BELL SYSTEM VALUE OF THE MARGINAL PRODUCT OF
 CAPITAL VS CAPITAL UNIT COSTS



1971, for example, the value of the marginal product generated by an incremental dollar of net capital stock was 31¢, while the cost of a dollar of capital input was 9.2¢. These results suggest that the Bell System is undercapitalized, as shown explicitly in Table 5 by comparing the Bell System optimal capital-labor ratios and the actual ratios over the postwar years.¹¹

□ **Comparison with the Bell System *M-NH* results.** The *A-NH* function was empirically tested on essentially the same aggregate Bell System data used by Vinod to test his *M-NH* formulation.¹² Comparisons of the *A-NH* results with the *M-NH* findings¹³ show that the two nonhomogeneous formulations produce very different economic results, despite the fact that both of them exhibit reasonably good statistical properties. The main differences follow. (1) The *M-NH* marginal product of labor is negative for the early postwar years and increases almost at as fast a rate as the marginal product of

TABLE 5
 BELL SYSTEM OPTIMAL VS.
 ACTUAL CAPITAL-LABOR
 RATIOS
 (ADDITIVE NONHOMOGENEOUS
 PRODUCTION FUNCTION)

YEAR	OPTIMAL CAPITAL/ LABOR RATIO	ACTUAL CAPITAL/ LABOR RATIO
	x_2/x_1 OPT.	x_2/x_1
1947	1.114	13.065
1948	7.971	14.496
1949	13.374	15.661
1950	15.871	16.661
1951	19.621	16.396
1952	26.714	17.175
1953	32.966	18.022
1954	37.631	19.155
1955	40.757	20.933
1956	48.805	20.861
1957	51.670	23.073
1958	54.377	24.913
1959	53.596	27.523
1960	56.513	28.985
1961	58.698	30.945
1962	61.116	32.860
1963	63.913	34.739
1964	67.785	35.771
1965	71.832	37.276
1966	76.695	37.429
1967	82.133	38.906
1968	90.175	39.637
1969	104.461	39.760
1970	124.918	40.561
1971	151.401	43.180

¹¹ It can be readily shown that the optimal capital-labor ratio, $\left(\frac{x_2}{x_1}\right)_{\text{opt.}}$, can be expressed as:

$$\left(\frac{x_2}{x_1}\right)_{\text{opt}} = \frac{\epsilon_2 W}{\epsilon_1 i}.$$

¹² The data used here for the *A-NH* formulation were updated for 1971 and include several very minor revisions.

¹³ See [6], p. 540.

capital, whereas, the *A-NH* marginal product of labor is positive for all observed periods and increases at a much slower rate than the marginal product of capital. (2) The *M-NH* marginal elasticity of labor rose at a faster rate than the marginal elasticity of capital, whereas the levels of the *A-NH* marginal elasticities of labor declined over the observed period, as the *A-NH* marginal elasticities of capital increased rapidly. (3) Analysis of the labor market based on the *M-NH* results suggests that for the period 1965 to 1970 the Bell System value of the marginal product of labor exceeded by far the compensation per man-hour. The *M-NH* findings indicate, for example, that the value of the marginal product of labor in 1970 was substantially above \$11, compared with a compensation per man-hour of \$5.4. The *A-NH* findings suggest, as discussed above, that the Bell System was extremely close to equilibrium in that market. (4) The *M-NH* results indicate diminishing returns to capital, while the *A-NH* finding indicate somewhat increasing returns to the capital input. The *A-NH* scale elasticity series shows higher levels of economies of scale prior to 1960 than the *M-NH* series. (5) The *A-NH* scale elasticities also show level fluctuations over the observed period, while the *M-NH* marginal elasticity series is monotonically increasing. The trend growth of scale elasticity is more moderate for the *A-NH* than for the *M-NH* functions. (6) The *M-NH* elasticities of factor substitution are positive for all observed periods and continuously increasing in value, the *A-NH* elasticities of factor substitution are negative for the first five years and continuously decreasing during the subsequent period. (6) The *M-NH* findings suggest convergence of the Bell System actual capital-labor ratios to the optimal ones, while the *A-NH* results indicate divergence, although both suggest undercapitalization.

The last point of difference requires some more elaboration. It should be borne in mind that the Bell System production pattern exhibits substantial economies of scale and technology over the observed period. Therefore, a strong deviation from the normally optimal may be, as pointed out before, partly due to a misspecification of technology. The essential difference between both formulations is that *A-NH* findings suggest that the Bell System shows close to optimal behavior in the labor market and increasing divergence in the capital market, resulting in increasing divergence of the optimal capital-labor ratios from the actual ones, while the *M-NH* findings suggest nonoptimal behavior in both factor markets, with strong divergence in the capital market, but also with increasing opposite divergence in the labor market. The latter partially offsets the capital divergence effect on the optimal capital-labor ratio, causing it to converge to the actual capital-labor ratio.

It should be noted at this point of the discussion that the *A-NH* results, suggesting increasing undercapitalization on the part of the Bell System during the postwar period, are not necessarily at variance with the Averch-Johnson effect. The A-J incentive to overcapitalize will materialize only in periods when the regulatory allowed rate of return exceeds the firm's perceived cost of capital.¹⁴ However, in many postwar years the opportunity cost of capital perceived by the Bell System was either higher or extremely close to the rate of return

permitted by regulation. In such periods, the pattern of the firm's actual input decisions relative to the optimal is *a priori* indeterminate.¹⁵ It seems to us, therefore, that any conclusions suggesting that our undercapitalization results violate theoretical expectations should await further analysis.

Reflecting on some of the above differences, we think that the *A-NH* formulation produces more plausible economic indicators with regard to Bell System market behavior as well as with regard to the levels and variations of Bell System's economies of scale than the *M-NH* results. We are aware, however, that it may be argued that the *M-NH* produces more credible economic findings with regard to deviations from the optimal capital-labor ratio path. We do not necessarily agree, but these differences serve to point out that even the relatively more sophisticated nonhomogeneous production function formulations, let alone the conventional ones, are highly experimental. A number of such formulations tested on the same data can produce equally good statistical fits, while generating diverse economic results.

□ **Western Electric *A-NH* production function of sealed contacts.**

The *A-NH* functional formulation was applied to data related to Western Electric Company's manufacture of electronic devices called sealed contacts. These data are available for the monthly time series of output y , and three inputs x_1 = labor, x_2 = capital, and x_3 = engineering, on the basis of 59 observations. The *A-NH* formulation is readily extendible to three (and more) inputs:

$$y = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_1 \ln x_2 + a_5x_1 \ln x_3 + a_6x_2 \ln x_1 + a_7x_2 \ln x_3 + a_8x_3 \ln x_1 + a_9x_3 \ln x_2. \quad (12)$$

The marginal products are easily derived from (12):

$$f_1 = \frac{\partial y}{\partial x_1} = a_1 + a_4 \ln x_2 + a_5 \ln x_3 + a_6 \frac{x_2}{x_1} + a_8 \frac{x_3}{x_1} \quad (13)$$

$$f_2 = \frac{\partial y}{\partial x_2} = a_2 + a_6 \ln x_1 + a_7 \ln x_3 + a_4 \frac{x_1}{x_2} + a_9 \frac{x_3}{x_2} \quad (14)$$

$$f_3 = \frac{\partial y}{\partial x_3} = a_3 + a_5 \ln x_1 + a_9 \ln x_2 + a_8 \frac{x_1}{x_3} + a_7 \frac{x_2}{x_3}. \quad (15)$$

Applying ordinary least squares to (12), we get:

$$\begin{aligned} y = & 0.0066x_1 - 11.2838x_2 + 8.4393x_3 \\ & (1.6914) \quad (-3.9197) \quad (2.7634) \\ & + 0.0036x_1 \ln x_2 + 0.0109x_1 \ln x_3 + 2.1439x_2 \ln x_1 \\ & (0.7488) \quad (1.2736) \quad (3.4769) \\ & + 5.6294x_2 \ln x_3 - 2.1579x_3 \ln x_1 - 2.5867x_3 \ln x_2 \\ & (7.0759) \quad (-3.3309) \quad (-5.4008) \\ R^2 = & 0.9761 \quad S.E. = 0.3774 \quad DW = 1.1070. \end{aligned} \quad (16)$$

The marginal elasticities of the three inputs and the scale elasticities are computed by applying the estimates in (16) to expressions

¹⁵ For a detailed discussion see [3], pp. 173-176.

(13), (14), and (15), and multiplying them by $\frac{x_1}{y}$, $\frac{x_2}{y}$, and $\frac{x_3}{y}$ series, respectively. These results for the 59 monthly periods are shown in Table 6.

TABLE 6

WESTERN ELECTRIC'S MARGINAL ELASTICITIES IN PRODUCTION OF SEALED CONTRACTS

MONTH	MARGINAL ELASTICITY OF LABOR (1)	MARGINAL ELASTICITY OF CAPITAL (2)	MARGINAL ELASTICITY OF ENGINEERING (3)	SCALE ELASTICITY (4) = (1) + (2) + (3)	MONTH	MARGINAL ELASTICITY OF LABOR (1)	MARGINAL ELASTICITY OF CAPITAL (2)	MARGINAL ELASTICITY OF ENGINEERING (3)	SCALE ELASTICITY (4) = (1) + (2) + (3)
1	-76.838	-78.240	161.944	6.866	31	1.180	0.791	6.148	8.119
2	-37.905	-29.978	80.952	13.089	32	0.498	-3.469	14.772	11.801
3	-25.269	-11.157	43.506	7.080	33	0.517	-1.922	10.572	9.166
4	-11.360	-5.749	20.668	3.560	34	1.770	2.417	4.612	8.799
5	-12.600	-11.292	28.818	4.926	35	1.484	1.899	4.507	7.890
6	-7.681	-5.989	16.753	3.082	36	1.528	1.949	4.955	8.433
7	-3.486	-10.599	16.991	2.906	37	1.470	2.057	4.547	8.074
8	-2.287	-6.690	10.994	2.017	38	1.292	1.966	3.826	7.083
9	1.280	-1.264	4.482	4.498	39	1.220	1.185	5.840	8.245
10	2.071	0.119	3.316	5.506	40	1.207	0.098	8.733	10.037
11	2.310	0.663	2.576	5.550	41	0.967	-0.371	8.819	9.416
12	2.067	0.332	2.619	5.019	42	0.920	-0.297	8.776	9.399
13	2.107	0.553	2.242	4.902	43	0.950	0.891	5.956	7.796
14	1.986	0.721	1.562	4.269	44	0.940	1.669	3.315	5.924
15	1.408	-0.006	1.929	3.331	45	0.931	1.331	4.849	7.111
16	1.093	-0.342	1.567	2.318	46	0.880	1.693	3.162	5.735
17	0.836	-0.452	1.637	2.021	47	1.047	2.270	0.244	3.561
18	0.589	-0.573	1.531	1.546	48	1.004	2.054	0.142	3.200
19	0.259	-1.126	2.107	1.240	49	0.943	2.117	0.340	3.400
20	0.008	-1.432	2.404	0.979	50	0.989	2.317	0.505	3.810
21	0.484	-0.597	1.334	1.221	51	0.937	2.244	0.606	3.787
22	1.098	0.151	0.555	1.804	52	0.792	1.974	0.667	3.433
23	1.729	1.754	2.429	5.911	53	0.766	1.709	0.240	2.715
24	1.612	1.680	2.314	5.606	54	0.743	1.675	0.263	2.680
25	1.668	1.930	1.618	5.216	55	0.737	1.798	0.382	2.917
26	1.959	2.558	2.187	6.704	56	0.698	1.693	0.360	2.751
27	1.962	2.647	1.812	6.421	57	0.759	1.798	0.341	2.898
28	1.614	2.213	1.901	5.728	58	0.761	1.827	0.360	2.947
29	1.412	1.209	5.421	8.042	59	0.812	2.135	0.798	3.745
30	1.101	0.034	7.782	8.917					

The elasticity series in Table 6 indicates some interesting economic trends. The levels of the marginal elasticities of engineering dropped dramatically below the elasticity of the other inputs during the 47 to 59 monthly periods. For most other observed periods, the elasticities of engineering stayed at higher levels than those of other inputs, with particularly high levels for the periods 1 to 10 and 29 to 46. Labor elasticities were generally stable except for the 1 to 8 period, when they fluctuated in the negative range. Capital elasticities maintained an average higher level during the 47 to 59 period than in previous periods. They were in the negative range for the periods 1 to 9, 15 to 21, 32 to 33, 41 to 42. Of the elasticities of all inputs, the marginal elasticity of engineering was the only one that never exhibited a negative value throughout the observed period. The scale elasticity exhibited a variable pattern, with relatively high levels during the 1 to 15 and 23 to 46 periods, and relatively low levels during the 16 to 22 and 46 to 59 periods. The major components

affecting the scale elasticity variations were the fluctuations in the marginal elasticity of engineering. The scale elasticity remained substantially higher than unity for all observed periods, except for the 20th, indicating impressive economies of scale throughout the observed production process.

□ **Comparison with *M-NH* results.** The *M-NH* formulation was tested by Vinod on the same data.¹⁶ Statistically, the *M-NH* produces significant estimates at the 5-percent level for all estimated parameters, while in the *A-NH* case, three of the estimated parameters (a_1 , a_4 , and a_5) are statistically not significant at the 5-percent level. Thus, the *M-NH* appears to generate a more stable formulation for the purposes of structural analysis, when applied to the same sample of Western Electric data.

Some of the similarities in *A-NH*'s and *M-NH*'s economic results for Western Electric relate to economies of scale, and the behavior of the marginal elasticities of labor and engineering in the late periods. Both the *A-NH* and *M-NH* scale elasticities series indicate impressive economies of scale throughout the period, with declines in the levels in recent periods relative to particularly high levels in the preceding periods. The *A-NH* and *M-NH* marginal labor elasticities remain relatively constant and close to unity throughout almost two-thirds of the observed periods (periods 24 to 59). The marginal elasticities of engineering for both formulations drop significantly in comparison with their preceding high levels. Otherwise, there are numerous differences in the levels and patterns of the *A-NH* and *M-NH* elasticities series, readily observed by comparing the results in Table 6 with Vinod's charts in Figures 1 and 2.¹⁷

□ **Bell Canada's *A-NH* aggregate production function.** The *A-NH* and *M-NH* formulations were also tested on Bell Canada's aggregate production data for the period 1951 to 1967.¹⁸ The *A-NH* formulation yields the following statistical results:

$$y = 23.5951x_1 - 0.8266x_1 - 3.6153x_1 \ln x_2 + 0.3010x_2 \ln x_1$$

$$(9.4227) \quad (-5.8345) \quad (-8.8924) \quad (8.0995) \quad (17)$$

$$R^2 = 0.9980 \quad S.E. = 7.0264 \quad DW = 1.2166,$$

where y = revenues in constant 1967 dollars, x_1 = man-hours, and x_2 = net capital stock.

The *M-NH* formulation produces the following estimates:

$$\ln y = 29.4280 - 7.3210 \ln x_1$$

$$(3.6227) \quad (-3.6031)$$

$$-3.5944 \ln x_2 + 1.1018 \ln x_1 \ln x_2 \quad (18)$$

$$(-2.9375) \quad (3.6033)$$

$$R^2 = 0.9927 \quad S.E. = 0.0338 \quad DW = 0.6103.$$

By comparing the statistical results in (17) and (18), it can be ob-

¹⁶ See [6], pp. 537-539.

¹⁷ See [6], pp. 538-539.

¹⁸ Data published in [4], p. 198.

TABLE 7

BELL CANADA MARGINAL PRODUCTS AND ELASTICITIES, 1952 TO 1967
(ADDITIVE NONHOMOGENEOUS PRODUCTION FUNCTION)

YEAR	MARGINAL PRODUCT OF LABOR	MARGINAL PRODUCT OF CAPITAL	MARGINAL ELASTICITY OF LABOR	MARGINAL ELASTICITY OF CAPITAL	SCALE ELASTICITY	ELASTICITY OF FACTOR SUBSTITUTION
	f_1	f_2	ϵ_1	ϵ_2	ϵ	σ
1952	\$4 475	\$0.031	1.007	0.087	1.094	-0.120
1953	4.427	0.057	0.939	0.165	1.104	-0.278
1954	4.346	0.086	0.878	0.252	1.130	-0.576
1955	4.138	0.121	0.808	0.363	1.171	-1.636
1956	3.969	0.155	0.745	0.475	1.221	56 911
1957	3.925	0.188	0.700	0.596	1.296	1.899
1958	4.144	0.215	0.697	0.731	1.428	1.078
1959	4.588	0.233	0.700	0.821	1.520	0.848
1960	5 227	0.244	0.724	0.895	1 619	0.745
1961	5.914	0.250	0.743	0.940	1.683	0.695
1962	6.519	0.257	0.732	0.947	1.678	0.663
1963	6.866	0.269	0.733	1.002	1.734	0.636
1964	7.095	0.281	0.721	1.043	1.763	0.615
1965	7.329	0.291	0.696	1.053	1.749	0.601
1966	7.642	0.301	0.675	1.061	1.736	0.589
1967	7 970	0.311	0.660	1.075	1.735	0.579

served that the *A-NH*, tested on Canadian data, generates substantially higher *t*-statistics than the *M-NH*. This suggests that the *A-NH* formulation may exhibit more stability for the purposes of structural analysis of Bell Canada data. The economic results of the *A-NH* and *M-NH* are shown in Tables 7, 8, and 9, respectively. Interestingly, the differences in economic interpretations are strikingly similar to the differences for the Bell System aggregate data as discussed above. This is self-evident upon a comparative examination of Tables 7 and 8.

TABLE 8

BELL CANADA PRODUCT ACCELERATIONS, 1952 TO 1967
(ADDITIVE NONHOMOGENEOUS FUNCTION)

YEAR	DIRECT ACCELERATION (LABOR)	DIRECT ACCELERATION (CAPITAL)	CROSS ACCELERATION (LABOR - CAPITAL)
	f_{11}	f_{22}	f_{12}
1952	\$ -0.843x10 ⁻¹	\$ 0.509x10 ⁻³	\$ 0.304x10 ⁻³
1953	-0.891x10 ⁻¹	0.429x10 ⁻³	0.749x10 ⁻³
1954	-0.924x10 ⁻¹	0.361x10 ⁻³	0.111x10 ⁻²
1955	-0.909x10 ⁻¹	0.300x10 ⁻³	0.129x10 ⁻²
1956	-0.905x10 ⁻¹	0.248x10 ⁻³	0.147x10 ⁻²
1957	-0.939x10 ⁻¹	0.199x10 ⁻³	0.172x10 ⁻²
1958	-0.105	0.153x10 ⁻³	0.211x10 ⁻²
1959	-0.122	0.119x10 ⁻³	0.254x10 ⁻²
1960	-0.145	0.932x10 ⁻³	0.298x10 ⁻²
1961	-0.168	0.756x10 ⁻⁴	0.336x10 ⁻²
1962	-0.187	0.635x10 ⁻⁴	0.362x10 ⁻²
1963	-0.196	0.551x10 ⁻⁴	0.368x10 ⁻²
1964	-0.200	0.491x10 ⁻⁴	0.367x10 ⁻²
1965	-0.204	0.443x10 ⁻⁴	0.366x10 ⁻²
1966	-0.210	0.397x10 ⁻⁴	0.367x10 ⁻²
1967	-0.215	0.355x10 ⁻⁴	0.367x10 ⁻²

TABLE 9

BELL CANADA MARGINAL PRODUCTS AND ELASTICITIES, 1952 TO 1967
(MULTIPLICATIVE NONHOMOGENEOUS PRODUCTION FUNCTION)

YEAR	MARGINAL PRODUCT OF LABOR	MARGINAL PRODUCT OF CAPITAL	MARGINAL ELASTICITY OF LABOR	MARGINAL ELASTICITY OF CAPITAL	SCALE ELASTICITY	ELASTICITY OF SUBSTITUTION
	f_1	f_2	ϵ_1	ϵ_2	ϵ	σ
1952	\$-1.543	\$0.211	-0.347	0.598	0.251	0.102
1953	-1.145	0.216	-0.243	0.620	0.378	0.146
1954	-0.631	0.224	-0.128	0.658	0.531	0.194
1955	0.064	0.245	-0.012	0.737	0.749	0.254
1956	0.817	0.264	-0.153	0.810	0.963	0.304
1957	1.692	0.272	-0.302	0.863	1.165	0.346
1958	2.694	0.258	-0.453	0.877	1.330	0.376
1959	3.825	0.244	-0.583	0.860	1.443	0.396
1960	5.072	0.225	-0.702	0.826	1.528	0.409
1961	6.378	0.211	-0.801	0.794	1.595	0.420
1962	7.921	0.211	-0.889	0.777	1.666	0.430
1963	9.142	0.213	-0.976	0.796	1.772	0.446
1964	10.376	0.222	1.054	0.824	1.878	0.460
1965	11.828	0.235	1.123	0.848	1.971	0.472
1966	13.501	0.246	1.193	0.867	2.061	0.483
1967	15.268	0.257	1.265	0.888	2.153	0.494

■ The *A-NH* function, tested on several sets of telecommunications data, produces, in general, reasonably good statistical results as well as plausible economic findings. In the case of the aggregate Bell System and Bell Canada Data the *A-NH* findings describe, in our opinion, a more realistic economic labor market behavior of the two companies than Vinod's *M-NH* formulation. Generally speaking, both functional forms suggest that the telecommunications production processes examined are nonhomogeneous and of variable elasticity of substitution, and exhibit impressive economies of scale which are variable over the scale of production. Otherwise, the *A-NH* and *M-NH* findings often imply quite diverse economic patterns of behavior for the same enterprises. Thus, the major point underscored by these comparative analyses is that production function formulations even as nonrestrictive as the nonhomogeneous forms discussed in this paper, are highly experimental. Two formulations, like the *A-NH* and *M-NH*, with almost equally attractive economic properties can, when tested on the same bodies of data, produce statistical results of similar quality, but radically diverse, although plausible, economic findings. Findings based on the more restrictive conventional forms such as Cobb-Douglas, *CES*, and *CRES* should be treated even with greater degree of caution unless they are tested successfully on several samples and exhibit a reasonable degree of stability. All this does not mean that findings generated by production functions cannot be potentially very useful, but it does serve to point out that production model estimates in telecommunications should be analyzed with extreme care, and solid conclusions with regard to the nature of the production process based on results of one particular model tested on one sample of a particular set of data should be avoided.

4. Concluding remarks

References

1. ALLEN, R. G. D. *Mathematical Analysis for Economists*. London: Macmillan & Co., 1956.
2. AVERCH, H. AND JOHNSON, L. L. "Behavior of the Firm under Regulatory Constraint." *The American Economic Review*, Vol. 52, No. 5 (December 1962), pp. 1053-1069.
3. BAUMOL, W. J. AND KLEVORICK, A. K. "Input Choices and Rate-of-Return Regulation: an Overview of the Discussion." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 162-190.
4. DOBELL, A. R., TAYLOR, L. D., WAVERMAN, L., LIU, T. H., AND COPELAND, M. D. G. "Telephone Communications in Canada: Demand, Production, and Investment Decisions." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 2 (Spring 1972), pp. 175-219.
5. FRISCH, R. *Theory of Production*. Chicago: Rand McNally & Co., 1965.
6. VINOD, H. D., "Nonhomogeneous Production Functions and Applications to Telecommunications." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 2 (Autumn 1972), pp. 531-543.

Some aspects of optimal pricing for telecommunications

Lyn Squire

World Bank
Washington, D. C.

The published work on telecommunications has been primarily concerned with marginal-cost pricing and the peak-load problem: externalities generated by the telephone have been ignored. This article reverses the situation by stressing two externalities—the one created when someone makes a call and the other created when a new subscriber joins the system. The first externality represents the benefit of an incoming call to the recipient of that call, a benefit for which he does not pay. The second arises from the fact that existing subscribers do not pay for the installation of the new subscriber's phone, but do obtain a benefit because they can now call the new subscriber if they so wish. Optimal prices are obtained allowing for these externalities and a measure is developed for the total benefits of the system.

■ The postwar literature abounds with articles on public utility pricing,¹ and recently there has been a spate of articles on external economies.² As yet, however, these various strands of thought have not been applied to telecommunications, which, perhaps of all public utilities, affords one of the best instances of external economies. In fact, it would appear that an externality is created every time a phone is used because the recipient of the call obtains a benefit for which he does not pay. Similarly, a new subscriber confers a benefit on existing subscribers because they are now able to call the new subscriber. It is these externalities which we shall examine here.

Thus far, the published literature on telecommunications is sparse. The best known article is that by Hazlewood,³ who adopts a pragmatic approach to the problem based on marginal-cost pricing. His main purpose is first to identify the relationships between different subscribers and different types of calls and their respective costs and then to price accordingly. Unallocable costs are to be distributed "in such a way as least to restrict the development of the service."⁴ This article set the tempo for ensuing publications.

Lyn Squire is an economist at the World Bank. He holds the B.Sc. (Econ.) from the University of Wales (1967), and the M.Sc. (Econ.) from the London School of Economics (1968), and is currently completing his Ph.D. thesis for submission to Cambridge University, England. His research interests are in share pricing for projects in underdeveloped countries.

The views expressed here are those of the author and do not necessarily reflect current thinking in the World Bank. The author, however, does wish to thank his colleagues at the Bank and the editor of this *Journal* for helpful comments.

¹ See [8] for examples.

² See [6] for a survey.

³ See [3].

⁴ See [3], p. 68.

1. Introduction

Shepherd considered the implications of peak-load pricing and concluded that "in Great Britain, residence telephone service is overpriced, relative to its cost, as compared with business service" because, although all business calls occur in the peak period, many residence calls occur in the nonpeak period.⁵ Finally, Littlechild applied a mathematical programming technique to the peak-load problem in telecommunications which enabled him to include both joint-demand and joint-cost characteristics.⁶

Typically, authors advocating marginal-cost pricing qualify their conclusions by observing that externalities have been excluded by assumption and that the pricing rules would have to be modified, if externalities did exist and were important. Neither Shepherd nor Littlechild makes such a qualification; Hazlewood, while making the qualification, does not develop the argument. In this article, we shall reverse the process, making some simple assumptions about cost and concentrating on the external economies.

2. The basic model

■ Throughout the analysis we shall employ two convenient simplifying assumptions: the first is that the demand for telephone calls does not fluctuate; the second is that the life of the telephone is very short. The first enables us to circumvent the peak-load pricing problem while the second allows us to dispense with a period analysis, and thereby avoid introducing a discount rate. We shall also take it to imply that installation charges and rentals become one and the same thing.

In order to focus the analysis directly on the externalities by reducing the complexity of the equations, we shall also assume that long-run marginal costs for both telephone installations and telephone calls are constant. Thus, the total cost (TC) of the system can be expressed as

$$TC = cZ + kN, \quad (1)$$

where

c = Cost per call,

Z = Number of calls,

k = Cost per installation, and

N = Number of subscribers.

Total benefits derive in part from making calls and in part from receiving calls. The n th subscriber's benefit from making calls can be measured by the area under his demand curve.⁷ This demand curve, however, will be determined to some extent by the number of subscribers because a call cannot be made without both the caller and the "callee's" possessing a phone. Thus, the n th subscriber's demand for calls will be a function of both the price of a call *and* the number of subscribers. Conceptually, we can perform an experiment in which we vary the price paid by the n th subscriber for making calls while we keep the size of the system constant. The area under

⁵ See [7], p. 351.

⁶ See [5].

⁷ The use of consumer surplus as a measure of welfare is, of course, subject to various qualifications. For a comprehensive discussion of consumer surplus and its limitations, see [2].

the resulting demand curve will represent the subscriber's willingness to pay for calls, which is our measure of his benefit from those calls. This experiment can be repeated individually for each of the N subscribers and the market demand curve can then be obtained by (horizontally) summing the individual demand curves.

However, we cannot actually perform this experiment; nor can we observe the resulting demand curve—which we shall therefore describe as the “conceptual” demand curve. If we try to perform the experiment in practice by increasing price to all subscribers, there will be a reduction in the demand for calls *and* in the number of subscribers both directly (because the price of calls has increased) and indirectly (because subscribers will anticipate fewer incoming calls).⁸ It follows that the observed market demand curve will be more elastic than the “conceptual” demand curve, which is the relevant demand curve for assessing consumer willingness-to-pay.⁹ Figure 1 brings out the relationship between these two different types of demand curve. The curves marked $\bar{N}_1, \bar{N}_2$, etc., show the conceptual demand curves for given system sizes. Each conceptual demand curve is intersected once by the observed demand curve. Thus, the observed demand curve delineates the single observed point on each conceptual demand curve (e.g., points A, B, C). We may note that the price elasticities of the conceptual and of the observed demand curves are related as follows: the price elasticity of the conceptual demand curve equals the price elasticity of the observed demand curve less the product of the elasticities of subscribers with respect to price and of calls with respect to subscribers.¹⁰

We can now write the benefit ($B_n^{\bar{N}}$) derived by the n th subscriber from outgoing calls as

$$B_n^{\bar{N}} = \int_0^{Z_n} g_n(Z_n, \bar{N}) dZ_n, \quad (2)$$

where Z_n is the number of calls made by the n th subscriber and g_n is his conceptual demand curve, where the bar over the N serves to remind us that the conceptual demand curve is derived with the system size held constant. Summing over the N subscribers to obtain

⁸ This conclusion is probably intuitively obvious, but it is brought out formally in equation (6), which shows that for the marginal subscriber the private benefit of phone ownership will just equal the private cost.

⁹ Leibenstein faced a similar problem in his analysis of bandwagon goods. In his case, however, the number of consumers was an argument in the *utility* function because consumers derived utility from the fact that other people were consuming that commodity. In our case, subscribers only derive utility from making (and receiving) calls, but the number of calls they can make is to some extent determined by the size of the system. Thus, whereas in Leibenstein's problem the number of consumers appeared as an argument in both the utility and the demand function, in our problem it only appears in the demand function. See [4].

¹⁰ The relationship between the elasticities is derived as follows:

$$\frac{dZ}{dp} (N = \text{const.}) = \frac{dZ}{dp} - \frac{dN}{dp} \frac{dZ}{dN},$$

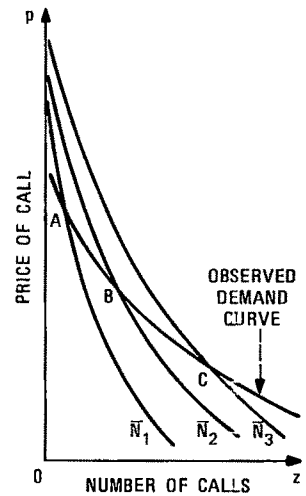
where p = the price per call, or

$$e = \zeta - \eta E,$$

where e = price elasticity on the conceptual demand curve, ζ = observed price elasticity, and η and E are, respectively, the elasticity of subscribers with respect to price and the elasticity of calls with respect to subscribers.

FIGURE 1

RELATIONSHIP BETWEEN THE CONCEPTUAL AND OBSERVED MARKET DEMAND CURVE FOR CALLS



the total benefit of making calls gives

$$\sum^N B_n \bar{N} = \sum^N \int_0^{Z_n} g_n(Z_n, \bar{N}) dZ = \int_0^Z G(Z, \bar{N}) dZ, \quad (3)$$

where $G(Z, \bar{N})$ is the market *conceptual* demand curve. For future reference, we can define the *marginal* (i.e., the N th) subscriber's benefit from outgoing calls as

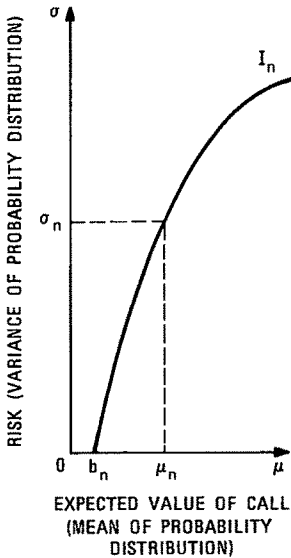
$$B_N \bar{N} = \sum^N B_n \bar{N} - \sum^{N-1} B_n \bar{N}. \quad (4)$$

Let us now turn to the benefit of an incoming call. When one's telephone rings, one has no idea of the possible benefit of the incoming call. In fact, in the case of nuisance calls, the benefit is presumably negative. We can, however, measure the benefit of incoming calls by asking the subscriber how much he would be willing to pay to be allowed to answer his phone. In other words, we are looking for the certainty equivalent of an uncertain return. Situations involving risk can be analyzed in terms of an indifference map relating risk and expected return. Usually, it is hypothesized that the possible return or benefit can be represented by a probability distribution, so that the expected return is measured by the mean of the distribution, and the risk by the variance. Figure 2 shows an indifference curve (I_n) relating risk and expected return (benefit) for the n th subscriber. Let the variance and mean for the n th subscriber's probability distribution for the benefit to be expected from an incoming call be σ_n and μ_n , respectively. From Figure 2 the subscriber is indifferent between paying b_n to answer his phone and not answering his phone. We may interpret b_n as the benefit from an incoming call.¹¹

We can deduce something about the probable value of b_n from observed behavior. In the first place, most subscribers do answer their phones, which operation presumably involves some private cost in terms of effort expended. (This private cost must be quite high in the case of the subscriber who gets out of his shower to answer his phone!) In some cases, the private cost may be negligible, but in general we can assume that it is not negative (i.e., subscribers do not derive utility from lifting the receiver), from which observation we can deduce that $b_n \geq 0$. This is not to say that b_n will never be negative. The subscriber who receives a *prolonged* series of nuisance calls learns not to answer his phone, and in extreme instances may even have his phone disconnected. Such behavior indicates that $b_n < 0$ for this subscriber. Alternatively, the subscriber may request an unlisted number which may be interpreted as an attempt to insure that b_n is positive, because the unlisted number will only be given to people from whom the subscriber wishes to receive calls.

Unfortunately, there is little else that we can deduce about the value of b_n , because there is no market mechanism whereby subscribers can express their preferences relating to incoming calls. The

FIGURE 2
CERTAINTY-EQUIVALENT OF AN
INCOMING CALL FOR A RISK-
AVERTER



¹¹ Figure 2 is drawn on the usual assumption that the individual is a risk averter. Note that we have not distinguished between *ex ante* and *ex post* utility. We can circumvent this problem by assuming that the probability distribution of the value of incoming calls is known with absolute certainty by the subscriber. The analysis also assumes that incoming calls are independent in the statistical sense.

only exception is the firm that offers a toll-free number. Such a firm must value incoming calls as much as, or more than, the price it has to pay for these calls. Throughout this article, as a reasonable first approximation, we shall assume that $b_n = b$ for all n (i.e., all subscribers have the same indifference map and the same probability distribution describing the benefit of incoming calls). Accordingly, we can write the total benefits (TB) derived from both making and receiving calls in this system as

$$TB = \sum^N B_n^{\bar{N}} + bY = \int_0^Z G(Z, \bar{N})dZ + bY, \quad (5)$$

where Y is the total number of incoming calls and equals Z , since each incoming call for the callee is an outgoing call for the caller.

The last equation that we need before the optimal pricing rules can be determined is the one governing the size of system. This equation depicts the condition that the marginal subscriber will equate the private benefit and cost of phone ownership. Specifically, for the marginal subscriber the benefit of making calls [$B_N^{\bar{N}}$] plus the anticipated benefit of receiving calls (bY_N) will just equal the cost of making calls (pZ_N) plus the rental, i.e.,

$$B_N^{\bar{N}} + bY_N = pZ_N + r, \quad (6)$$

where p is the price per call, r is the rent, and Y_N is the number of incoming calls anticipated by the N th subscriber. Thus, other things being equal, the number of subscribers will diminish if p is increased or if Y_N is reduced. Note that $B_N^{\bar{N}} - pZ_N$ may be interpreted as the consumer surplus on outgoing calls accruing to the marginal subscriber. For future use, we shall define consumer surplus accruing to the n th subscriber in a system of N subscribers as

$$S_n^{\bar{N}} = B_n^{\bar{N}} - pZ_n. \quad (7)$$

We are now in a position to derive the optimal price and rent for this system. We want to choose p and r such that $TB - TC$ is a maximum [i.e., we want to maximize the difference between equations (5) and (1)]. Recalling that TB was derived from the conceptual demand curve (i.e., with N constant), we can differentiate $TB - TC$ with respect to Z to obtain the optimal price given a particular system size. Setting the derivative equal to zero, we obtain

$$p^* = c - b, \quad (8)$$

where p^* is the optimal price. Marginal-cost pricing would, of course, imply $p^* = c$, which means that price would be set too high by a margin which represents the external benefit of an incoming call.

We can derive an optimal price for each system size.¹² The optimal price and the system size will in turn determine the optimal benefits for that system size through equation (5) and the total costs through equation (1). Furthermore, given the price of calls, the rent, which is necessary to ensure that size of system, is uniquely determined by

¹² In fact, with our particular cost assumptions p^* is independent of the size of system, but in practice this is unlikely to be so if returns to scale tend not to be constant. On this point and with regard to other aspects of the analysis, compare the literature on the two-part tariff.

equation (6). The optimal rent is then determined by expanding the system size until net benefits are maximized, the benefits of each system size being calculated with the price of calls set at its optimum for that system size.

For example, the increase in net benefits resulting from an expansion of the system by one subscriber may be written as

$$\Delta(TB - TC) = \left[\sum_{n=N}^N B_n^{\bar{N}} - \sum_{n=N-1}^{N-1} B_n^{\bar{N}-1} \right] + b(Z_N + Y_N) - c(Z_N + Y_N) - k. \quad (9)$$

To the right of the equality sign, the first term is the additional benefit of making more calls, the second term is the additional benefit of receiving more calls (Z_N —the number of *outgoing* calls made by the new subscriber—are received by existing subscribers and Y_N by the new subscriber from existing subscribers), the third term is the additional cost of those calls, and the last term the cost of the new installation. Making use of equations (4) and (6) and setting $\Delta(TB - TC)$ equal to zero to obtain the maximum, equation (9) becomes

$$r^* + p^*Z_N^* = k + c(Z_N^* + Y_N^*) - \left[\sum_{n=N-1}^{N-1} B_n^{\bar{N}} - \sum_{n=N-1}^{N-1} B_n^{\bar{N}-1} \right] - bZ_N^*, \quad (10)$$

which displays the standard result that in the presence of externalities price should be set such that the private benefit ($r^* + p^*Z_N^*$) equals the social cost [$k + c(Z_N^* + Y_N^*)$] less the externality which in this case comprises (1) $\left[\sum_{n=N-1}^{N-1} B_n^{\bar{N}} - \sum_{n=N-1}^{N-1} B_n^{\bar{N}-1} \right]$, the benefit to existing subscribers of making Y_N^* calls to the new subscriber,¹³ and (2) bZ_N^* , the benefit to existing subscribers of receiving Z_N^* calls from the new subscriber.

The expression derived for the optimal price [equation (8)] can now be substituted directly into equation (10). It is instructive, however if we go back one step and make the substitution in equation (9). Setting $\Delta(TB - TC)$ equal to zero and making use of our definition of consumer surplus [equation (7)], we obtain

$$\left[\sum_{n=N-1}^{N-1} S_n^{\bar{N}} - \sum_{n=N-1}^{N-1} S_n^{\bar{N}-1} \right] + S_N^{\bar{N}} = k, \quad (11)$$

which says that

$$\Delta \text{ in consumer surplus to existing subscribers} + \text{consumer surplus to the new subscriber} = \text{cost of installation}.$$

In the event that the marginal cost of calls is not constant, this result would have to be adjusted to include changes in producer surplus,

¹³ Raul Schkolnick has pointed out to me that existing subscribers may receive a benefit from a new subscriber, even if they do *not* call him. For example, if the new subscriber is a doctor, his *availability* may be considered a benefit even if there is never any actual need to call him. However, one would imagine that the marginal benefit from an extra doctor would decrease quite rapidly, so that it is perhaps not too misleading to ignore this externality. On the other hand, new subscribers probably enjoy a benefit from the fact that they are now able to call a doctor should they require one. But this will be reflected in the amount of rent a subscriber is willing to pay and is not, therefore, an externality.

but the adjustment is relatively minor. The optimal rent can now be derived by substituting equation (6) into equation (11):

$$r^* = k + bY_N^* - [\sum_{n=1}^{N-1} S_n \bar{N} - \sum_{n=1}^{N-1} S_n \bar{N}^{-1}]. \tag{12}$$

Thus, the optimal rent will be greater or less than marginal cost, depending on the relative strength of two opposing forces. The benefits received by the marginal subscriber from incoming calls (bY_N^*) are a reason for having the optimal rent greater than marginal cost, whereas the external benefit enjoyed by existing subscribers $-[\sum_{n=1}^{N-1} S_n \bar{N} - \sum_{n=1}^{N-1} S_n \bar{N}^{-1}]$ —is a reason for having an optimal rent below marginal cost. The rent should only be set equal to marginal cost if the two opposing forces are of equal strength or if the new subscriber does *not* receive any calls because then both bY_N^* and $[\sum_{n=1}^{N-1} S_n \bar{N} - \sum_{n=1}^{N-1} S_n \bar{N}^{-1}]$ are zero. Note that if it is argued that the benefit from incoming calls is zero so that $p^* = c$, we still cannot equate rent to the marginal cost of installation. In fact, if $b = 0$, we know that r^* should be less than k by a margin which measures the consumer surplus accruing to existing subscribers from making calls to the new subscriber. On the other hand, it is not inconceivable that $b > c$, which would imply a negative optimal price. The effect of such an outcome on the optimal rent is not clear: bY_N^* would be large because b is large, but the additional consumer surplus enjoyed by existing subscribers would also be large because price is negative. Note, however, that we do not need a prior estimate of b in order to deduce the optimal pricing rules. Provided we know how the conceptual demand curve moves as system size increases, we can immediately determine the optimal price from equation (11), because the optimal price determines the consumer surplus accruing to existing subscribers and the new subscriber. In other words, we can use the equation for the optimal price [equation (8)] *to determine b given p^** , rather than *to determine p^* given b* . Then, given an estimate of the number of incoming calls (Y_N^*) expected by the marginal subscriber, r^* can be determined from equation (12).¹⁴

¹⁴ The following numerical examples are intended simply to convey the flavor of the theoretical results. For convenience assume that the total consumer surplus generated by a new subscriber is divided equally between existing subscribers and the new subscriber so that $r^* = k/2 + bY_N^*$. If we assume values of b expressed as a proportion of c , we can immediately obtain the ratio of p^* to c from equation (8). To obtain r^* as a ratio of k , we need an estimate of c/k and Y_N^* . We shall assume that $c/k = 0.02$ (i.e., installation costs fifty times as much as a call) and take values of $Y_N^* = 25$ and 100:

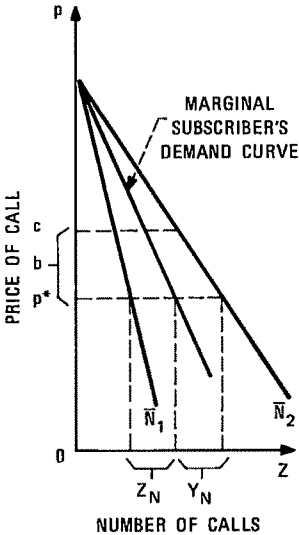
b/c	p^*/c	$c/k = 0.02$	
		$Y_N^* = 25$	$Y_N^* = 100$
0.000	1.000	0.500	0.500
0.250	0.750	0.625	1.000
0.500	0.500	0.750	1.500
1.000	0.000	1.000	2.500
2.000	-1.000	1.500	4.500.

The results support the statements made in the text (i.e., the larger b , the smaller p^* and the larger r^* ; and the larger Y_N^* , the larger r^*).

3. Implications for investment

■ Two questions are of obvious importance with respect to investment decisions given this pricing system: (1) can we measure the social benefits of the system? and (2) how do we finance any deficits which may be implied by our pricing rules? To clarify the answers to these questions it is useful to specify the form of the conceptual demand curve as the system size increases. Let us assume that the conceptual demand curves are straight lines and that the price elasticity *for a given price* remains constant as the system size increases. This implies that the conceptual demand curve for each size of system has the same vertical intercept but a different slope. Note that this condition does not impose any restriction on the observed demand curve, because no restriction has been imposed on the change in the slope of the conceptual demand curve as the system size increases. We shall also assume that this constant elasticity condition applies to the conceptual demand curve of the marginal subscriber in each system size.¹⁵ Finally, assume that the number of incoming calls received by the marginal subscriber equals the number of his outgoing calls so that $Z_N = Y_N$.

FIGURE 3
ILLUSTRATIVE EXAMPLE



These assumptions are reflected in the construction of Figure 3. The addition of one subscriber changes the conceptual demand curve from N_1 to N_2 . The marginal subscriber's demand curve for calls apportions the total increase in calls between those originating with the marginal subscriber and those originating with other subscribers. Note that the assumptions imply that consumer surplus is *proportional* to the number of calls made, so that given $Y_N = Z_N$

$$S_N \bar{N} = \frac{k}{2} \quad (13)$$

and

$$\sum^N S_n \bar{N} = \frac{k}{2} \frac{Z}{Z_N}. \quad (14)$$

We can now examine the questions posed at the start of this section in terms of relationships between total benefits (TB), total cost (TC), and total revenue (TR). Using the definitions of TB and TC [equations (5) and (1), respectively] and the expression for the optimal price [equation (8)], we obtain

$$TB - TC = \sum^N S_n \bar{N} - kN^* \quad (15)$$

and from equation (14)

$$TB - TC = kN^* \left[\frac{1 - E^*}{E^*} \right], \quad (16)$$

where $E^* = 2Z_N^* N^* / Z^*$. Because $2Z_N^*$ is the total number of calls generated by the N th subscriber, E^* may be interpreted as the elasticity of calls with respect to subscribers, the asterisk serving to remind us that the elasticity is measured for the optimal pricing system. From equation (16) it follows that for a socially profitable system we want $E^* \leq 1$, which means that the increase in calls (both

¹⁵ This is not necessarily implied by the previous assumption, because the movement from one conceptual demand curve to the next involves an increase in calls emanating in part from the marginal subscriber, but also in part from existing subscribers.

incoming and outgoing) associated with each new subscriber *decreases* with the size of the system. This would seem to be a reasonable expectation.

The equation for TR is

$$TR = r^*N^* + p^*Z^*. \quad (17)$$

Recalling the expression for r^* [equation (12)] and making use of equation (13), we obtain

$$TC - TR = \frac{k}{2}N^* + bZ^*\left[\frac{2 - E^*}{2}\right], \quad (18)$$

which implies

$$TB - TR = r^*N^*\left[\frac{2 - E^*}{E^*}\right]. \quad (19)$$

Equations (18) and (19) tell us that $TC > TR$ and $TB \geq TR$ if $E^* \leq 2$. Thus, we see that if the system is socially profitable (i.e., $E^* \leq 1$), the enterprise necessarily runs a commercial deficit, and total revenue understates total benefits. If, for example, we assume that $E^* = 1$, so that the enterprise is a marginally acceptable investment, then both TB and TC exceed TR by r^*N^* .¹⁶

If an enterprise has a deficit, economists usually recommend some form of differential pricing which does not affect the marginal conditions. Thus, with telephones we could charge a higher price for intramarginal calls than for marginal calls or, we could charge a higher rent for intramarginal subscribers than for marginal subscribers. The first method, however, would affect the decision to own a phone because the benefit of having a phone includes the benefit from making calls *less* their price.

The second method is economically acceptable only if the income elasticity of the demand for calls is zero. And, of course, any system of differential pricing is usually politically unacceptable and is anyway difficult to administer.

A preferable solution would be to charge subscribers for incoming calls. In this way, we could arrange for the total price (paid in part by the caller and in part by the callee) to equal just marginal cost, and with marginal cost equal to average cost the enterprise would break even on the provision of calls. This pricing system, however, will affect the optimal rent. Setting the charge for incoming calls at b and for outgoing calls at p^* (equal to $c - b$), equation (6) must now be rewritten as

$$r = B_N\bar{N} - p^*Z_N = S_N\bar{N}$$

¹⁶ If we abandon the specific assumptions determining the conceptual demand curves, then a socially profitable system requires $\sum_N S_N\bar{N} > kN^*$ which, using equation (11), implies that the *average* consumer surplus must exceed the *marginal* consumer surplus (i.e., the total consumer surplus generated by the marginal subscriber). $TC > TR$ and $TB > TR$ will occur if $kN^* > r^*N^* - bY^*$, which, using equation (12), requires

$$N^*bY_N^* - bY^* - N^*\left[\sum_{N=1}^{N-1} S_N\bar{N} - \sum_{N=1}^{N-1} S_N\bar{N-1}\right] < 0.$$

Clearly, this condition will be fulfilled if $Y_N^* < Y^*/N^*$ (i.e., if the number of incoming calls per subscriber decreases as the system size increases. Given the specific assumptions in the main text, this last condition was implied by the condition for a socially profitable system.

and this gives

$$r^* = \frac{k}{2}. \quad (20)$$

This pricing system will result in an increase in revenue provided $E < 2$, which we know to be the case in a socially profitable system.¹⁷ However, we still have a deficit given by

$$TC - TR = \frac{kN^*}{2}. \quad (21)$$

This deficit arises because existing subscribers receive a benefit for which they do not pay when a new subscriber joins the system. Unlike the external benefit derived from an incoming call, it is difficult to see how any *simple* pricing system could capture the external benefit accruing to existing subscribers from calls made to the new subscriber. Theoretically, each of the existing subscribers who calls the new subscriber should have his rent increased by the consumer surplus derived from those calls. It is difficult to envisage such a system's ever being used, which means that telecommunications enterprises must either abandon the optimal pricing rules or else run a deficit.

4. Concluding remarks

■ We have demonstrated in a simple telecommunications system that pricing according to marginal cost is not economically efficient. In general, the price of calls should be less than their marginal cost by a margin representing the external benefit received by the "callee." Rents, however, should not necessarily be less than the marginal cost of installing a phone. If the benefit derived by the marginal subscriber from receiving calls exceeds the consumer surplus accruing to existing subscribers from making those calls, then the rent should exceed marginal cost and *vice versa*.

Moreover, if benefits are approximated by revenue, then investment in telecommunications will be suboptimal. This, of course, is a standard result, because approximating benefits by revenue means that the consumer surplus is excluded. Note, however, that in the case of telecommunications the rent captures the consumer surplus accruing to the *marginal* subscriber, so that revenue only understates benefits by the *difference* between the consumer surpluses enjoyed by the *average* and *marginal* subscribers times the number of subscribers. In other words, other things being equal, revenue is not so bad a measure of benefits in telecommunications as it may be in other industries which only produce one output and therefore only set one price.

At the theoretical level it would be interesting to remove some of the restrictive assumptions. Adopting a multiperiod analysis, introducing a peak demand, and assuming that returns to scale are not constant would certainly complicate the optimal pricing rules, but one suspects that the nature of the results would be unchanged. At the empirical level it would be useful to investigate the relationship

¹⁷ This result is intuitively obvious because the benefit from incoming calls was only captured by the first pricing system through the inclusion in the rent of the term bY_N . The second pricing system captures it directly so, if $Y > NY_N$ (or $E < 2$), the second pricing system will increase revenue.

between the observed and conceptual demand curves as the system size is increased, in terms of the relationship between their price elasticities. We could also obtain market information about the value of incoming calls by charging subscribers for such calls (although the author knows of no telephone company which does employ such a pricing system). In particular, it would be interesting to see if, say, business subscribers place a higher value on incoming calls than, say, residential subscribers. If this were the case, then the assumption that $b_n = b$ for all subscribers would have to be abandoned.

Other characteristics of telephone services could also be introduced into the analysis. Many telephone systems in the world today are congested in the sense that the probability of achieving a connection immediately is low. Recommending prices below marginal cost may, therefore, exacerbate an already inferior system. We may counter with two arguments: firstly, we could determine optimal prices which take explicit account of congestion as represented by a functional relationship between the number of calls and the delay per call; and secondly, we could analyze the optimal relationship between calls and delay. There is, however, a more fundamental problem which must be considered. The analysis set out to establish economically efficient pricing rules in a situation where externalities are prevalent. This meant, in general, that the enterprise ran at a commercial loss. Such a situation may be acceptable in developed countries but in the underdeveloped countries subsidies from general taxation are not easy to secure. It might be interesting, therefore, to reexamine the pricing rules, given the constraint that total revenue at least equals total cost. Alternatively, we could introduce directly into the analysis the premium that the government places on its own revenue.¹⁸ This latter method would, of course, secure an economically efficient result and could mean, for instance, that we use telecommunications as a source of general revenue.

References

1. ACHARYA, S. N. "Public Enterprise Pricing and Social Benefit-Cost Analysis." *Oxford Economic Papers*, Vol. 24, No. 1 (1972).
2. CURRIE, J. M., MURPHY, J. A., AND SCHMITZ, A. "The Concept of Economic Surplus and Its Use in Economic Analysis." *Economic Journal*, Vol. 81, No. 324 (1971).
3. HAZLEWOOD, A. "Optimum Pricing as Applied to the Telephone Service." *Review of Economic Studies*, Vol. 18(2), No. 46 (1950-1951).
4. LEIBENSTEIN, H. "Bandwagon, Snob and Veblen Effects in the Theory of Consumers' Demand." *Quarterly Journal of Economics*, Vol. 64, No. 2 (1950).
5. LITTLECHILD, S. C. 1970, "Peak-Load Pricing of Telephone Calls." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 191-210.
6. MISHAN, E. J. "The Postwar Literature on Externalities: an Interpretive Essay." *Journal of Economic Literature*, Vol. 9, No. 1 (March 1971).
7. SHEPHERD, W. G. "Residence Expansion in the British Telephone Service." *Journal of Industrial Economics*, Vol. 14, No. 3 (1966).
8. TURVEY, R. *Public Enterprise*. Harmondsworth, England: Penguin Modern Economics, 1968.

¹⁸ See [1].

Consumption externalities and imperfect corrective pricing

Peter A. Diamond

Professor of Economics
Massachusetts Institute of Technology

Pricing congested facilities above marginal production cost is a conventional approach to improving resource allocation. Where everyone is producing the same externality, a uniform price (in excess of marginal cost by the value of the externality) permits the competitive equilibrium to be Pareto optimal. Where individuals give rise to different externalities, but a uniform price is in effect, we have a second-best situation. When demands depend only on price, price should exceed marginal cost by a weighted average of externalities generated, the weights being the price derivatives of demand. When demands also depend on congestion, the optimal price generally diverges from this rule. The price should be lower when the individuals giving large external diseconomies per unit demanded tend to be price insensitive and congestion sensitive in their demands (relative to the average). In this case public expenditures to decrease congestion directly should not be carried to the point where the marginal direct benefit from congestion reduction equals the marginal cost. Optimal income distribution is also examined.

1. Introduction

■ In the presence of consumption externalities, a natural first response to improve resource allocation is corrective pricing (i.e., prices different from marginal production cost or, equivalently, the imposition of excise taxes). The formal modeling supporting this policy recommendation shows the possibility of achieving a Pareto optimal allocation by levying corrective surcharges (price premiums or taxes) under competitive conditions.¹ The models have considered two different situations²—(1) price additives which differ across individuals and require only convexity of preferences of each individual to achieve Pareto optimality³ and (2) consumption externalities

Peter A. Diamond received the A.B. degree from Yale University in 1960 and the Ph.D. from the Massachusetts Institute of Technology in 1963. At M.I.T. Professor Diamond's research centers on public finance.

I wish to thank A. Friedlaender, J. Mirrlees, and R. Solow for helpful comments and the National Science Foundation for financial support of this research.

¹ This proposition should not be confused with its converse, which is not generally valid—the presence of excise taxes equal to the value of external diseconomies caused is not sufficient to ensure Pareto optimality. This is a familiar problem of multiple equilibria with taxation (local optima which are not global optima) which we would expect to arise in many externality situations.

² These models consider only externalities that are related to market behavior and do not consider changes in the use of commodities which alter externalities. Thus these models are not comparing corrective taxes to antilitter advertising. Artificially one could consider discarded litter as another commodity, but this is only a serious alternative for some easily measurable types of pollution.

³ See, e.g., Starrett [9].

◀which depend only on aggregate demand and are thus correctable
■by uniform taxation.⁴ In most real-world situations, however,
■government-imposed surcharges cannot vary from transaction to
■transaction. In this paper I shall examine the role of corrective pricing
■to improve efficiency when externalities vary with the individuals
◀causing them, but prices are uniform.

A tax per gallon on leaded gasoline which does not vary with the age or weight of the car in which it is used, and so does not vary with the pollution emitted per gallon is one example of a uniform corrective surcharge. Similarly, another example would be a peaking-period price additive on a congested bridge, which does not vary with fast and slow drivers who contribute differently to congestion. More generally, government pricing above marginal production cost when enjoyment of the facilities depends on the demands of others is an example of corrective pricing. When the enjoyment varies with the particular other people using the facility but the price is uniform, we have corrective pricing which cannot achieve Pareto optimality. If this correction is introduced by a government-owned, government-regulated, or just public-spirited firm to increase social welfare, it is natural to call it a congestion surcharge. The obvious modification of the standard theory would be a surcharge equal to some average of the externalities caused. However this modification ignores the interactions of the demands of the users of the facility when the externalities affect demands as well as utilities.⁵ This interaction can move the optimal surcharge outside the range of externalities at the equilibrium, so it is unequal to any internal average. Even the widely valid public finance proposition that some corrective taxation raises welfare may fail to be true.

The paper starts with the simple case where externalities do not affect demands. In this case the optimal surcharge is a weighted average of externalities, the weights being the responses of demand to price increases. Dropping the separability assumption, but preserving one of its implications, that all demands decrease with the surcharge across equilibria, the optimal surcharge remains a weighted average of externalities, the weights being the responses of demands to price increases across equilibria. However, without the separability assumption it is possible, and often plausible, that some demands will rise with the surcharge across equilibria (because their increase in response to the decreased demand of others exceeds the direct effect of price on demand). When some demands increase and others decrease, the weighted average of externalities involves both positive and negative weights, and thus it need not lie in the range of externalities. This is illustrated by a two-consumer example, where both consumers give diseconomies but the optimal surcharge is zero. (This is not an extreme example in the sense that subsidizing diseconomies may be optimal.)

To get a better feeling for the range of solutions, further analysis is confined to the case where externalities work through a single ag-

⁴ See, e.g., Aoki [1], Chipman [3], Malinvaud [7], and Samuelson [8]. As with these models, Pareto optimality is examined in this paper. The problems of an imperfect distribution of income are ignored.

⁵ This interaction is also a possible source of counterintuitive quantity differences between the competitive equilibrium and the full optimum. See Buchanan and Kafoglis [2] and Diamond and Mirrlees [5].

gregate, called congestion, for concreteness. Congestion is assumed to depend on individual demands and on government expenditures. (Congestion may be additive in demands but the weights must not be the same if the problem is to be preserved.) Optimality rules for surcharge and expenditure generally differ from the simple weighted average rule in the separable case for the surcharge and from the comparison of direct benefits with costs for expenditures. The magnitude of the difference depends on the covariance of marginal externalities with price derivatives of demand compared with the covariance of marginal externalities with congestion derivatives of demand. When price increases tend to discourage demanders giving small externalities, while the induced congestion decreases encourage demanders giving large externalities, both corrective tax and expenditure tend to be small at the optimum. Analysis with respect to this point is made under the assumption of utilities linear in income available for other consumption. In Section 6, this assumption is dropped to examine income redistribution which affects externalities. The marginal utilities of income of the different individuals at the optimum are related to their contributions to congestion per unit demanded and their income derivatives of demand.

2. Separable externalities

■ We shall start with the case which leads to a simple intuitive solution—where utilities are linear in income and separable in externalities.⁶ Let us denote by α_h the quantity demanded by individual h of the good being analyzed. Denote by μ_h and m_h his income available for all other goods and total income. We shall write his utility as

$$U^h(\alpha_1, \alpha_2, \dots, \alpha_n) + \mu_h$$

and assume that U^h is concave in α_h and twice continuously differentiable. We shall assume that all externalities are diseconomies

$$\frac{\partial U^h}{\partial \alpha_i} \leq 0 \quad \text{for } i \neq h \quad (1)$$

and that the marginal utility of own consumption is independent of the demands of others

$$\frac{\partial^2 U^h}{\partial \alpha_h \partial \alpha_i} = 0 \quad \text{for } i \neq h. \quad (2)$$

Given these assumptions, the demand of an individual does not depend on his income (we ignore the constraint $\mu_h \geq 0$) or the demands of others and depends just on the price charged to him. We shall denote the (constant) production cost of this good by p and the price charged the consumer by $p + t$, with t being a tax imposed by the government or congestion surcharge added by the firm. Utility maximization, viewing the behavior of others as given, can be expressed as

$$\begin{aligned} & \max_{\alpha_h} U^h(\alpha_1, \alpha_2, \dots, \alpha_n) + \mu_h \\ & \text{subject to: } (p + t)\alpha_h + \mu_h = m_h. \end{aligned} \quad (3)$$

The first-order condition is thus

$$\frac{\partial U^h}{\partial \alpha_h} \leq p + t \quad \text{with} \quad \alpha_h \left(\frac{\partial U^h}{\partial \alpha_h} - p - t \right) = 0. \quad (4)$$

Since marginal utility depends only on the quantity demanded, we can write demand as a function of the price to the consumer, by solving (4):

$$\alpha_h^* = \alpha_h(p + t). \quad (5)$$

From the concavity assumption, demand is decreasing with price, $\alpha_h' \leq 0$. It is clear that a demand function which does not depend on the demand of others is inappropriate for congestion problems. However, in many pollution problems, the tendency to pollute will not vary with the aggregate level of pollution. Unless a firm's activity level depends on using river water, the effluents dumped into a river will not depend on the quality of the river water.

Let us now examine the relationship between the sum of utilities and the price charged to the consumer.⁷ The sum of utilities is

$$\sum U^h(\alpha_1, \dots, \alpha_n) + \sum \mu_h.$$

The constraint on the economy as a whole is that total income is divided between income available for other goods and the resource cost of producing the good being examined:

$$p \sum \alpha_h + \sum \mu_h = \sum m_h. \quad (6)$$

Viewing the constraint in this form, we assume that the revenue arising from sales at a price above production cost is returned to consumers as a lump sum. Since we are examining the sum of utilities and each utility is linear in income, the sum does not depend on the particular recipients of this revenue. Thus, the sum of utilities can be written as a function of the surcharge:

$$W(t) = \sum U^h[\alpha_1(p + t), \dots, \alpha_n(p + t)] - p \sum \alpha_h(p + t) + \sum m_h. \quad (7)$$

Differentiating with respect to the surcharge, we have

$$W'(t) = \sum_h \sum_i \frac{\partial U^h}{\partial \alpha_i} \alpha_i'(p + t) - p \sum_h \alpha_h'(p + t). \quad (8)$$

Making use of the first-order condition for individual choice, (4), we can express this as

$$W'(t) = \sum_h \sum_{i \neq h} \frac{\partial U^h}{\partial \alpha_i} \alpha_i' + t \sum_h \alpha_h'. \quad (9)$$

Since each demand function is nonincreasing with price and utilities are nonincreasing in demands of others, welfare increases with the surcharge at a zero rate [$W'(0) > 0$] provided some externality is present. To calculate the surcharge which maximizes the sum of

⁷ Given the linearity of utility in income available for other consumption, maximization of a weighted sum of utilities would imply zero levels of consumption of all other goods for some consumers. I ignore this uninteresting range of solutions.

utilities we set (9) equal to zero and solve for t :

$$t^* = \frac{- \sum_i \sum_{h \neq i} \frac{\partial U^h}{\partial \alpha_i} \alpha_i'}{\sum_i \alpha_i'} \quad (10)$$

Thus the optimal surcharge is a weighted average of the externalities, $\sum_{h \neq i} \frac{\partial U^h}{\partial \alpha_i}$, the weights being the price derivatives of demand, α_i' .

3. Decreasing demand

■ We now drop the assumption that the externalities are separable, (2). This implies that the first-order conditions for individual demand depend on the levels of demand of all other consumers:

$$\frac{\partial U^h(\alpha_1, \dots, \alpha_n)}{\partial \alpha_h} \leq p + t \quad \text{with} \quad \alpha_h \left(\frac{\partial U^h}{\partial \alpha_h} - p - t \right) = 0. \quad (11)$$

Let us now assume that these n equations have simultaneous non-negative solutions as a function of the price $\alpha_h(p + t)$, and that each equilibrium demand is decreasing in price, $\alpha_h' \leq 0$. We can now repeat the analysis above, replacing the demand functions of the individuals by the functions giving general equilibrium demands. All the conclusions continue to hold with this replacement.⁸ The first-order condition (10) is valid without the assumption of decreasing demands, but it does not convey much general information about the magnitude of the optimal surcharge, since the weights do not necessarily lie in the unit interval. The following two-person example illustrates the lack of general information, for the optimal surcharge will be zero despite the presence of diseconomies from the demands of each of the individuals.

Assume that the two utility functions take the particular forms:

$$\begin{aligned} U^1(\alpha_1, \alpha_2) + \mu_1 &= \alpha_1^{\frac{1}{2}} - \frac{1}{3}\alpha_2 + \mu_1 \\ U^2(\alpha_2, \alpha_2) + \mu_2 &= 0.3 \log(\alpha_2 + 0.9\alpha_2) - \alpha_1 + \mu_2. \end{aligned} \quad (12)$$

Given these utility functions, the demands are

$$\begin{aligned} \alpha_1 &= 0.25(p + t)^{-2} \\ \alpha_2 &= \text{Max}\{0.3(p + t)^{-1} - 0.9\alpha_1, 0\}. \end{aligned} \quad (13)$$

Across equilibria the demand of individual one and aggregate demand are both decreasing in price. However, for $p + t < 1.5$ the demand of individual two increases with price, preventing the application of the analysis above. Summing utilities net of production cost for $\alpha_2 > 0$, we have a welfare function:

$$\begin{aligned} W(t) &= m_1 + m_2 + 0.3 \log 0.3 - 0.3 \log(p + t) \\ &\quad + 0.1(p + t)^{-1} - 0.2(p + t)^{-2}. \end{aligned} \quad (14)$$

For $p = 1$, this function achieves its maximum⁹ at $t = 0$. The attempt

⁸ Clearly the result is unchanged if there exist individuals with increasing demands who cause no externalities.

⁹ One can check local optimality by calculating $W'(0) = 0$, $W''(0) = -0.7$. To confirm global optimality we note first that α_2 is positive for $t > -\frac{1}{2}$ and W

to introduce a surcharge decreases the first person's demand but increases the second person's. These offsetting effects happen to just balance at a zero surcharge. Of course, by altering the parameters it is possible to construct an example where it is optimal to subsidize.

Now let us make the simplifying assumption that the form of externality can be described by a single aggregate which is disliked by everyone. This will enable us to relate the optimal surcharge to demand characteristics rather than the more complicated equilibrium properties. Let us refer to it as "congestion" and denote it by γ :

$$\gamma = \gamma(\alpha_1, \dots, \alpha_n). \quad (15)$$

We shall denote the partial derivatives by γ_h ($\gamma_h \geq 0$). We now write utility as

$$U^h(\alpha_h, \gamma) + \mu_h.$$

With this form demand is a function of the price and level of congestion, $\alpha_h = \alpha_h(p + t, \gamma)$. Given concavity of U in α , α_h is non-increasing in price and decreasing where positive.

The equilibrium level of congestion is assumed to be well-defined and satisfies

$$\gamma = \gamma[\alpha_1(p + t, \gamma), \dots, \alpha_n(p + t, \gamma)]. \quad (16)$$

Thus the price derivative of the equilibrium level of congestion satisfies

$$\frac{d\gamma}{dt} = \frac{\sum \gamma_h \frac{\partial \alpha_h}{\partial t}}{1 - \sum \gamma_h \frac{\partial \alpha_h}{\partial \gamma}}. \quad (17)$$

We assume that this derivative is positive. This will clearly hold in the interesting case where congestion decreases demand.

We now face the question of describing individual demand; in particular, of deciding whether individuals pay attention to their own effect on congestion in determining individual demand.¹⁰ If individuals do consider this relation, the first-order condition becomes

$$\frac{\partial U^h}{\partial \alpha} + \gamma_h \frac{\partial U^h}{\partial \gamma} \leq p + t. \quad (18)$$

Given this behavioral assumption, the example above is consistent with an externality in terms of an aggregate. We merely define γ

has no other extrema in this range. For t between $-\frac{1}{4}$ and -1 , α_2 is zero and $W(t) = m_1 + m_2 + 0.3 \log 0.225 - 0.6 \log(p + t) + 0.5(p + t)^{-1} - 0.5(p + t)^{-2}$. Over this range $W'(t)$ is positive. Thus $W(t)$ is increasing up to $-\frac{1}{4}$ where the value coincides with that given by (14). One slightly unsatisfactory element of the example as it stands is that $\frac{\partial U^2}{\partial \alpha_1} < 0$, only for surcharges below $\frac{1}{4}$. However, this can be corrected by subtracting from U^2 as given a term which is positive and increasing in α_1 for α_1 below 0.2. This does not alter the optimality of the zero tax.

¹⁰ Differences in the answer to this question in a production context lie behind the differences in analysis between Goetz and Buchanan [6] and Aoki [17] and Chipman [3].

4. Externality aggregate

as $\alpha_1 + \alpha_2$ and replace α_2 by $\gamma - \alpha_1$ in U^1 ; and α_1 , by $\gamma - \alpha_2$ in U^2 . Thus we maintain the possibility of the curious results above. Therefore, even when γ is simply the sum of demands, uniform corrective taxation will not in general exactly restore efficiency in this case.¹¹

If we assume that individuals ignore their impact on congestion, the first-order conditions for individual demand become

$$\frac{\partial U^h}{\partial \alpha} \leq p + t. \quad (19)$$

It is now the case that there is an externality to everyone in the economy, and not just the other consumers. Thus if congestion is a simple sum of demands, everyone generates the same externality and so uniform corrective taxation does restore efficiency.¹² With congestion responding differently to the demands of different people (e.g., fast vs. slow drivers), uniform pricing cannot restore efficiency. However, the fact that congestion decreases with a surcharge ensures that the optimal surcharge is not zero, although it need not lie in the range of externalities. We now assume that individual demand satisfies (19),—i.e., individuals ignore their own impact on congestion. This is a natural assumption with a large number of consumers and is simple to work with. It does not differ in the limit from the alternative assumption, where the contribution of any one individual to congestion gets arbitrarily small as the number of individuals gets large.

We now seek to maximize social welfare $W(t)$, defined as

$$\sum U^h(\alpha_h, \gamma) + \sum m_h - p \sum \alpha_h.$$

As before, social welfare is unaffected by income distribution, given the linearity of utility in income with the same coefficients for everyone. Calculating the response of welfare to tax changes, we have

$$W'(t) = \sum \left(\frac{\partial U^h}{\partial \alpha} - p \right) \left(\frac{\partial \alpha_h}{\partial t} + \frac{\partial \alpha_h}{\partial \gamma} \frac{d\gamma}{dt} \right) + \frac{d\gamma}{dt} \sum \frac{\partial U^h}{\partial \gamma}. \quad (20)$$

From the first-order condition for individual choice we can rewrite this as

$$W'(t) = t \left(\sum \frac{\partial \alpha_h}{\partial t} + \frac{d\gamma}{dt} \sum \frac{\partial \alpha_h}{\partial \gamma} \right) + \frac{d\gamma}{dt} \sum \frac{\partial U^h}{\partial \gamma}. \quad (21)$$

Thus, at a zero level of surcharge, welfare increases with a surcharge if there is any externality present as congestion eases with the surcharge. Setting W' equal to zero and substituting for $\frac{d\gamma}{dt}$ from (17), we have the first-order condition for the optimal surcharge¹³:

¹¹ See also Goetz and Buchanan [6].

¹² See Aoki [1], Chipman [3], and Malinvaud [7].

¹³ Note that a constant value for γ_A makes the surcharge equal to the marginal externality, which is the same for everyone. If γ is the sum of individual demands, the Pareto optimal allocation can be achieved.

$$\begin{aligned}
& -\frac{d\gamma}{dt} \sum \frac{\partial U^h}{\partial \gamma} \\
& = \frac{\sum \frac{\partial \alpha_h}{\partial t} + \frac{d\gamma}{dt} \sum \frac{\partial \alpha_h}{\partial \gamma} - \sum \frac{\partial U^h}{\partial \gamma} \sum \gamma_h \frac{\partial \alpha_h}{\partial t}}{\sum \frac{\partial \alpha_h}{\partial t} - \left(\sum \gamma_h \frac{\partial \alpha_h}{\partial \gamma} \right) \left(\sum \frac{\partial \alpha_h}{\partial t} \right) + \left(\sum \gamma_h \frac{\partial \alpha_h}{\partial t} \right) \left(\sum \frac{\partial \alpha_h}{\partial \gamma} \right)} \quad (22) \\
& = -\sum \frac{\partial U^h}{\partial \gamma} \left[\frac{\sum \gamma_h \frac{\partial \alpha_h}{\partial t}}{\sum \frac{\partial \alpha_h}{\partial t}} \right] \\
& \quad \times \left[\frac{\sum \frac{\partial \alpha_h}{\partial t}}{\sum \frac{\partial \alpha_h}{\partial t} - \left(\sum \gamma_h \frac{\partial \alpha_h}{\partial \gamma} \right) \left(\sum \frac{\partial \alpha_h}{\partial t} \right) + \left(\sum \gamma_h \frac{\partial \alpha_h}{\partial t} \right) \left(\sum \frac{\partial \alpha_h}{\partial \gamma} \right)} \right].
\end{aligned}$$

The first term of the last expression is the value of a decrease in congestion. The second is the average change in congestion from an increase in individual demand. The third is an adjustment for interactions. If individual demands do not depend on congestion ($\frac{\partial \alpha_h}{\partial \gamma} = 0$), we have the separable case described above; the third term equals one and the optimal surcharge is a weighted average of externalities, the weights being the price derivatives of demand. To determine whether the surcharge is above the weighted average of externalities, we need to determine the sign of the difference term in the denominator above. Let us call it Δ :

$$\Delta = \sum \gamma_h \frac{\partial \alpha_h}{\partial t} \sum \frac{\partial \alpha_h}{\partial \gamma} - \sum \gamma_h \frac{\partial \alpha_h}{\partial \gamma} \sum \frac{\partial \alpha_h}{\partial t}. \quad (23)$$

Since $\sum \frac{\partial \alpha_h}{\partial t}$ is negative, the optimal surcharge is below (above) the weighted average of externalities when Δ is negative (positive and less than $-\sum \frac{\partial \alpha_h}{\partial t}$). Assuming that $\sum \frac{\partial \alpha_h}{\partial \gamma}$ is not zero, we can write

Δ as

$$\Delta = \left(\sum \frac{\partial \alpha_h}{\partial t} \right) \left(\sum \frac{\partial \alpha_h}{\partial \gamma} \right) \left[\sum_h \gamma_h \frac{\partial \alpha_h / \partial t}{\sum_i \partial \alpha_i / \partial t} - \sum_h \gamma_h \frac{\partial \alpha_h / \partial \gamma}{\sum_i \partial \alpha_i / \partial \gamma} \right]. \quad (24)$$

For the interesting case where $\sum \frac{\partial \alpha_h}{\partial \gamma}$ is negative, Δ is negative pro-

vided the covariance¹⁴ of γ_h with $\frac{\partial \alpha_h}{\partial \gamma} / \sum \frac{\partial \alpha_i}{\partial \gamma}$ is greater than the covariance of γ_h with $\frac{\partial \alpha_h}{\partial t} / \sum \frac{\partial \alpha_i}{\partial t}$. Thus the optimal surcharge will be small relative to the average externality when individuals who contribute greatly to congestion per unit demanded (large γ_h) tend to have demands which are congestion sensitive ($-\frac{\partial \alpha_h}{\partial \gamma}$ large) and price insensitive ($-\frac{\partial \alpha_h}{\partial t}$ small).

The absolute level of price derivatives of demand (given their distribution) has no effect on the optimal tax. In contrast, the greater the absolute level of congestion derivatives of demand (given their distribution), the further the tax moves away from the weighted average of externalities.

5. Public expenditures

■ The assumption thus far has been of a level of congestion determined solely by individual demands. Now let us add the possibility of a congestion decreasing expenditure, e . We would now rewrite the congestion function as¹⁵

$$\gamma = \gamma(\alpha_1, \dots, \alpha_n, e). \quad (25)$$

Differentiation of this relation gives the equilibrium response of congestion to expenditures:

$$\frac{d\gamma}{de} = \frac{\partial \gamma / \partial e}{1 - \sum \gamma_h \partial \alpha_h / \partial \gamma}. \quad (26)$$

The introduction of e does not alter the analysis of the optimal surcharge considered above. Subtracting the cost of expenditures, e , from the welfare function, we can calculate the first-order condition for optimal expenditures as

$$\frac{\partial W}{\partial e} - 1 = \sum \left[\left(\frac{\partial U^h}{\partial \alpha} - p \right) \frac{\partial \alpha_h}{\partial \gamma} + \frac{\partial U^h}{\partial \gamma} \right] \frac{d\gamma}{de} - 1 = 0. \quad (27)$$

Substituting from the individual first-order conditions, we can write this as

$$\left(t \sum \frac{\partial \alpha_h}{\partial \gamma} + \sum \frac{\partial U^h}{\partial \gamma} \right) \frac{d\gamma}{de} = 1. \quad (28)$$

Thus the equilibrium change in congestion is evaluated by the sum of revenue generated by the congestion surcharge and utility gained by the congestion decrease. Evaluating this at the optimal surcharge t^* , and substituting for t^* and $\frac{d\gamma}{de}$, we can express the first-order condi-

¹⁴ The remaining terms of the covariance expressions in the brackets sum to zero.

¹⁵ We assume $\frac{\partial \gamma}{\partial e} \leq 0$, $\frac{\partial^2 \gamma}{\partial e^2} \geq 0$.

on as

$$\frac{\partial \gamma}{\partial e} \sum \frac{\partial U^h}{\partial \gamma} = \frac{\Delta + \sum \frac{\partial \alpha_h}{\partial t}}{\sum \frac{\partial \alpha_h}{\partial t}}, \quad (29)$$

where Δ was defined above in (23). Production is only carried to a point where the marginal value of congestion decrease still exceeds one (the marginal cost) when Δ is negative. Thus, cases where the optimal surcharge is below the weighted average of externalities are cases where expenditures to ease congestion have a direct marginal value in excess of their marginal cost. A problem arises with widening a congested road (which has an optimal toll), not because demand increases to offset the congestion decrease, but because the demand increase is concentrated among users who are large contributors to congestion per trip relative to the weighted average of congestion generated per trip, the weights being the price derivatives of demand. Conversely expenditures should be pushed beyond the equality of direct benefits and costs when the induced demand gives rise to below average externalities. Given an optimal toll, inefficiency does not arise from too much demand in aggregate (since the toll controls the aggregate demand level), but from demand by the wrong individuals—too much by large externality producers and too little by small externality producers.

To look at the efficiency role of income redistribution let us drop the assumption of linearity of utility in income to allow income effects in the demand for the externality producing good.¹⁶ This will lead to a desired income distribution which may not equate social marginal utilities of income, with income tending to be redistributed to shift demand from large externality producers to small ones. We now write individual utility as

$$U^h(\alpha_h, \gamma, \mu_h).$$

We continue to assume that the first-order conditions ignore the indirect effect, through γ , of demand on utility giving the first-order conditions

$$\frac{\partial U^h}{\partial \alpha} \leq (p + t) \frac{\partial U^h}{\partial m} \quad \text{with} \quad \alpha_h \left(\frac{\partial U^h}{\partial \alpha} - (p + t) \frac{\partial U^h}{\partial m} \right) = 0, \quad (30)$$

where $\frac{\partial U^h}{\partial m}$ is the marginal utility of income. We assume that these n equations have solutions $\alpha_h(p + t, \gamma, m_h)$ —where m_h is total income—which are consistent with the definition of γ , and thus give γ as a function of the consumer price and the vector of incomes. We assume that demand is normal, implying

$$\frac{\partial \alpha_h}{\partial m} \geq 0, \quad \frac{\partial \alpha_h}{\partial t} \leq 0.$$

¹⁶ For policy discussions it would be silly to assume that income redistribution could vary with congestion propensities while the congestion surcharge could not. This analysis may be of interest in extending understanding of the interaction of different taxes in a second-best setting.

6. Income effects

For convenience we assume that the congestion surcharge revenue is returned to the individual purchaser as a lump sum. Given the ability to redistribute, this does not affect possible equilibria. The social welfare problem can be set up as

$$\text{Max}_{\gamma, t, m_1, m_2, \dots, m_n} \sum U^h(\alpha_h, \gamma, m_h - p\alpha_h) \quad (31)$$

$$\text{subject to: } \sum m_h = m, \quad \gamma = \gamma(\alpha_1, \dots, \alpha_n).$$

Forming a Lagrangian expression, we can write it as

$$L(\gamma, t, m_1, \dots, m_n, \lambda, \phi) = \sum U^h(\alpha_h, \gamma, m_h - p\alpha_h) - \lambda(\sum m_h - m) - \phi[\gamma(\alpha_1, \dots, \alpha_n) - \gamma], \quad (32)$$

where $\alpha_h = \alpha_h(p + t, \gamma, m_h)$. Differentiating first with respect to the surcharge, we have

$$\sum \left(\frac{\partial U^h}{\partial \alpha} - p \frac{\partial U^h}{\partial m} \right) \frac{\partial \alpha_h}{\partial t} - \phi \sum \gamma_h \frac{\partial \alpha_h}{\partial t} = 0. \quad (33)$$

Making use of the individual first-order conditions and solving for ϕ , we can express the Lagrangian as

$$\phi = \frac{t \sum \frac{\partial U^h}{\partial m} \frac{\partial \alpha_h}{\partial t}}{\sum \gamma_h \frac{\partial \alpha_h}{\partial t}}. \quad (34)$$

Thus ϕ is the ratio of the impacts caused by surcharge induced changes in demand on utility because of additional revenue to that on the congestion level. We can obtain an alternative expression for ϕ by differentiating the Lagrangian with respect to γ

$$\begin{aligned} & t \sum \frac{\partial U^h}{\partial m} \frac{\partial \alpha_h}{\partial \gamma} + \sum \frac{\partial U^h}{\partial \gamma} - \phi \left(\sum \gamma_h \frac{\partial \alpha_h}{\partial \gamma} - 1 \right) = 0 \\ \text{or} \quad & \phi = \frac{t \sum \frac{\partial U^h}{\partial m} \frac{\partial \alpha_h}{\partial \gamma} + \sum \frac{\partial U^h}{\partial \gamma}}{\sum \gamma_h \frac{\partial \alpha_h}{\partial \gamma} - 1}. \end{aligned} \quad (35)$$

The numerator of the expression is the impact of congestion change on social welfare, while the denominator has a multiplier appearance, reflecting the fact that a change in individual demands directly affecting congestion will have an indirect feedback on demands and, so, on congestion. Equating (34) and (35), we have a first-order condition without Lagrangian multipliers:

$$t^* = \frac{- \sum \frac{\partial U^h}{\partial \gamma} \sum \gamma_h \frac{\partial \alpha_h}{\partial t}}{\sum \frac{\partial U^h}{\partial m} \frac{\partial \alpha_h}{\partial t} \left(1 - \sum \gamma_h \frac{\partial \alpha_h}{\partial \gamma} \right) + \sum \frac{\partial U^h}{\partial m} \frac{\partial \alpha_h}{\partial \gamma} \sum \gamma_h \frac{\partial \alpha_h}{\partial t}}. \quad (36)$$

Of course, this reduces to (22) when all marginal utilities of income are unitary.

Turning now to the income distribution conditions we have, upon differentiation of L , the n equations

$$\frac{\partial U^h}{\partial m} \left(1 + t \frac{\partial \alpha_h}{\partial m} \right) - \lambda - \phi \gamma_h \frac{\partial \alpha_h}{\partial m} = 0. \quad (37)$$

Thus, variations in income derivatives of demand and in congestion impacts of demand generally result in unequal marginal utilities of income at the optimum. To explore who has low marginal utilities of income (and, *ceteris paribus*, high incomes), let us rewrite this condition as

$$\left(\lambda - \frac{\partial U^h}{\partial m} \right) = \gamma_h \frac{\partial \alpha_h}{\partial m} \left[t \frac{\frac{\partial U^h}{\partial m}}{\gamma_h} - \phi \right]. \quad (38)$$

Recalling from (34) that ϕ is a ratio of weighted averages for the entire economy of $t \frac{\partial U^h}{\partial m}$ and γ_h , a key element is whether the individual is above or below average in this ratio of effects of his demand. Given our sign assumptions, we have

$$\frac{\partial U^h}{\partial m} \begin{matrix} \geq \\ \leq \end{matrix} \lambda \quad \text{as} \quad t \frac{\frac{\partial U^h}{\partial m}}{\gamma_h} \begin{matrix} \geq \\ \leq \end{matrix} \phi, \quad (39)$$

or equivalently,

$$\frac{\partial U^h}{\partial m} \begin{matrix} \geq \\ \leq \end{matrix} \lambda \quad \text{as} \quad \gamma_h \begin{matrix} \geq \\ \leq \end{matrix} \frac{t\lambda}{\phi}.$$

Comparing (38) for two individuals with the same income derivatives of demand at the optimum, we see that the one having the greater marginal impact on congestion has the larger marginal utility of income (virtue rewarded). Comparing individuals giving the same marginal externality at the optimum, the sign of the effect of a larger income derivative of demand depends on whether marginal utility of income is above or below λ .

For $\frac{\partial U^h}{\partial m} \begin{matrix} \geq \\ \leq \end{matrix} \lambda$ the greater values of $\frac{\partial \alpha_h}{\partial m}$ are associated with

$$\begin{matrix} \text{larger} \\ \text{same} \\ \text{smaller} \end{matrix} \quad \text{values of } \frac{\partial U^h}{\partial m}. \quad (40)$$

Summarizing the terms, the population divides into heavy congesters per trip $\left(\gamma_h > \frac{t\lambda}{\phi} \right)$ and light congesters, with the latter having lower marginal utilities of income (and, *ceteris paribus*, higher incomes). Among heavy congesters, those with large income derivatives tend to have large marginal utilities of income (low incomes). Among light congesters, those with large income derivatives tend to have small marginal utilities of income (high incomes). Thus, depicting this crudely, income is transferred from large to small contributors

to congestion per trip. Among these small contributors, income is distributed to increase demand. Among the large contributors, it is distributed to decrease demand.

References

1. AOKI, M. "Marshallian External Economies and Optimal Tax-Subsidy Structure." *Econometrica*, Vol. 39, No. 1 (January 1971), pp. 35-54.
2. BUCHANAN, J. AND KAFOGLIS, M. "A Note on Public Goods Supply." *The American Economic Review*, Vol. 53, No. 3 (June 1963), pp. 403-414.
3. CHIPMAN, J. "External Economies of Scale and Competitive Equilibrium." *Quarterly Journal of Economics*, Vol. 84, No. 3 (August 1970), pp. 347-385.
4. DAVIS, O. AND WHINSTON, A. "Externalities, Welfare, and the Theory of Games." *Journal of Political Economy*, Vol. 70, No. 3 (June 1962), pp. 241-262.
5. DIAMOND, P. AND MIRRELES, J. "Aggregate Production with Consumption Externalities." *Quarterly Journal of Economics*, Vol. 87, No. 1 (February 1973), pp. 1-24.
6. GOETZ, C. AND BUCHANAN, J. "External Diseconomies in Competitive Supply." *The American Economic Review*, Vol. 61, No. 5 (December 1971), pp. 883-890.
7. MALINVAUD, E. *Leçons de Théorie Microéconomique*. Paris: Dunod, 1969.
8. SAMUELSON, P. "Pure Theory of Public Expenditure and Taxation," in *Public Economics*, J. Margolis and H. Guitton, eds., London: MacMillan & Co., 1969.
9. STARRETT, D. "Fundamental Nonconvexities in the Theory of Externalities." *Journal of Economic Theory*, Vol. 4, No. 2 (April 1972), pp. 180-199.

Market changes in the computer services industry

Richard S. Bower

Professor of Business Economics
The Amos Tuck School of Business Administration
Dartmouth College

Competition is not a natural state. Under it, firms become dissatisfied with performance and work to change the structure. What they seek is a structure that supports sufficient discipline to assure individual actions consistent with common interests. Often, the changes have been effected before the industry is prominent enough to command public attention. This is not so in computer services. The industry, already prominent, is still competitive. In this paper, the author describes the industry and suggests why and how changes in its structure may occur.

■ For those who find regulation an unsatisfactory alternative, even when industry conditions seem least likely to cause firms to act in the public interest, the decision of the Federal Communications Commission in its computer inquiry should be particularly satisfying. The Commission concluded “. . . that the offering of data processing services is essentially competitive and that . . . there is no public interest requirement for regulation by government of such activities.”¹ Our study of the computer services industry confirms this conclusion, but suggests that it soon may be open to challenge. The structure of the industry is vulnerable and will be changed for the worse if public policy actions are not carefully chosen. Poor choices could permit competition to deteriorate and revive regulation as an alternative.

Good choices require some understanding of the industry. As we define the computer services industry, it includes establishments that use computers as a principal input and that sell computer-generated information as their principal product, establishments often described as computer service bureaus, data banks, time-sharing vendors, on-line business services, and remote batch sellers. In this paper our purpose is to describe this industry in a way that makes clear why its structure is both satisfactory and vulnerable and to provide some indication of which policy choices are the good ones.

Richard S. Bower received the A.B. degree from Kenyon College (1949), the M.B.A. from Columbia University (1955), and the Ph.D. from Cornell University (1962). He is Professor of Business Economics at The Amos Tuck School of Business Administration at Dartmouth College. His current research interest is the regulation of health care delivery.

The study from which this paper was developed was supported by the Brookings Institution. Both substance and form of this paper benefitted from work by David Downes, Ivy Bosson, and Dorothy Bower. A legion of Tuck School students contributed to the study as did many people in the industry. In spite of all this help, the paper would never have been completed if Paul MacAvoy had not acted as head coach and first cheerleader.

¹ See [70], p. 8.

1. Introduction

2. The industry

■ Although economists have been concerned about problems of a "computer utility" for some time² and visions of the wide range of dial-up information services have been described,³ the industry is still unfamiliar and vaguely pictured. To introduce it we take the buyer's point of view.

Business firms, nonprofit institutions, government agencies, and other organizations require access to, and machine processing of internal and external data. The data, processing power, storage capacity, applications programs, and communications channels to do these things are all available from firms in the computer services industry. (1) Conventional service bureaus offer customers transactions processing performed locally by using their own computer programs and batch-processing techniques.⁴ In processing a payroll, for example, information supplied by the customer is punched at the service bureau, placed in a queue, run to completion on the computer with a proprietary program package, and returned to the customer and his employees in the form of summary reports and payroll checks. (2) On-line dedicated applications firms also offer transactions processing but differ from conventional service bureaus in gathering information over telephone lines which connect the customer directly with the computer. The recording of savings deposits and the up-dating of pass books is a familiar on-line application. (3) Data bank reference systems operate either off- or on-line, but unlike firms in the two sectors already described, they provide external information to customers rather than process internal information. Credit reference is a common service, but a better suggestion of the opportunities for applications is provided by the National Association of Securities Dealers Automatic Quotation System which makes available to brokers immediately updated bid-asked offers from market makers in over-the-counter stocks. (4) Firms in the remote job entry and (5) time-sharing sectors not only make some data banks and applications programs available to customers but also maintain simple languages that permit customers to write their own programs. The customers of firms in both these sectors have access to the computer directly from remote terminals via telephone lines. Although these sectors may be distinguished in terms of computer processing technique, limitations on program size, and the immediacy of system response, the distinctions are rapidly disappearing. A simulation model of the customer and his industry or a program to provide desirability indicators for an investment project is suggestive of what customers might create and run on the computer systems of either of these groups of suppliers.

These sectors make up an "industry." Firms in these sectors, much more than firms outside them, come close to making available to an organization its information requirements. They do not offer computer hardware, consulting, software, and raw data separately for the customer to assemble but, instead, combine them in a package that is ready for application. Their services are distinctive as far as

² See, for example, [4] or [30].

³ See [21].

⁴ Conventional service bureaus also sell computer time or raw power, which gives the customer use of the computer and freedom to use his own programs. These sales amount to only 10 percent [27], p. A-2, or 15 percent [37], p. 35, of total sales.

“demanders” are concerned. For smaller firms the speed, accuracy, and manipulative power associated with information from the computer services industry differ greatly from what is available outside the industry. The same is true for particular types of transactions and certain data banks. But there is also distinctiveness as far as suppliers are concerned. Although entry into the computer services industry is easy, particularly for those already selling aids for in-house computing, it is very difficult for firms outside the five sectors to respond to price increases within a sector by switching their operations to comparable services. Higher prices for a system that combines up-to-date bond prices, customer portfolio files, and an algorithm for switching bonds in the portfolio will be affected very little by software houses turning to the production of switching algorithms. They will be forced down only if the software houses combine these computer programs with data banks and hardware and make the full information package available; that is, only if they enter the computer services industry.⁵

Estimates of 1971 revenues for the industry, its sectors, and some benchmark activities are presented in Table 1.⁶ The conventional

TABLE 1
1971 REVENUE ESTIMATES (MILLIONS OF DOLLARS)

	ESTIMATES ¹	
	LOW	HIGH
CONVENTIONAL SERVICE BUREAUS	650 ^q	1040 ⁱ
DATA BANK REFERENCE SYSTEMS	115 ⁱ	150 ^a
ON-LINE DEDICATED APPLICATIONS	20 ^a	55 ⁱ
REMOTE JOB ENTRY	100 ⁱ	220 ^a
TIME-SHARING	185 ⁱ	220 ^a
<u>COMPUTER SERVICES INDUSTRY²</u>	1080 ^q	1500 ^{i,a}
REVENUES OF SOFTWARE SELLERS	450 ⁱ	650 ^a
REVENUES OF FACILITIES MANAGERS	300 ^a	650 ^a
TOTAL ELECTRONIC DATA PROCESSING EXPENDITURES FOR SALARIES, EQUIPMENT RENTALS AND OTHER REQUIRED INPUTS IN THE UNITED STATES ³	12500 ^q	22600

1 THE SOURCE OF THE ESTIMATE IS INDICATED BY a FOR ARTHUR D. LITTLE, INC [37], i FOR INTERNATIONAL DATA CORPORATION [27] AND q FOR QUANTUM SCIENCE CORPORATION [51].

2 THE COMPUTER SERVICES INDUSTRY LOW AND HIGH ESTIMATES ARE NOT THE SUMS OF THE HIGHS AND LOWS FOR THE INDUSTRY'S FIVE SECTORS. THEY ARE BASED ON FIVE SECTOR TOTALS FROM EACH SOURCE.

3 THE 22.6 BILLION HIGH ESTIMATE IS A ROUGH CALCULATION OF OUR OWN. INTERNATIONAL DATA CORPORATION [28], P. 6, REPORTS THAT \$26.4 BILLION OF COMPUTERS MADE BY AMERICAN-BASED COMPANIES AND VALUED AT ORIGINAL PURCHASE PRICE WERE IN USE IN THE UNITED STATES IN 1971. MONTHLY RENTAL RATES ARE ESTIMATED TO BE 1/40 OF PURCHASE PRICE [36] AND ARE REMARKABLY STABLE OVER A COMPUTER SYSTEM'S LIFE SO THAT \$7.9 BILLION IS THE ESTIMATED ANNUAL RENTAL OF THE 1971 COMPUTER STOCK. TAKING THE DIEBOLD [17], P. 30, SURVEY RESULT THAT THE TYPICAL INDUSTRIAL CORPORATION SPENDS 35 PERCENT OF ITS ELECTRONIC DATA PROCESSING BUDGET ON HARDWARE THE \$22.6 BILLION TOTAL EXPENDITURE IS IMPLIED.

⁵ It is difficult to define satisfactorily an industry in a world where buyers' alternatives seem to approach a continuum and producers' facilities are quite versatile. The problems are discussed by Scherer [53], pp. 52-57, and we have become painfully aware of them. Our definition of the computer services industry is the same as that used by International Data Corporation [27], but is narrower than that used by Quantum Science Corporation [51] and Arthur D. Little [37]. The Quantum and ADL definitions both include software houses, facilities managers, and most of the group serving in-house computer operations. Selwyn [54] suggests that a proper industry definition should include all computer users, and

service bureaus sector, which provides services locally or off-line, has revenues that at least equal the combined revenues of the four on-line or remote sectors, all of which deliver their services over telephone lines to customers remote from their computers. The industry, although its revenues exceed those of software houses or facilities managers, has no more than a 12-percent share⁷ of the total expenditures for information generated with computers. Clearly, neither current size nor relative position helps to explain the policy interest in the industry and its on-line sectors.

The growth in revenues since 1961, which is shown in Table 2, contributes something to an explanation. Since 1966, when the on-line

TABLE 2

REVENUE ESTIMATES 1961-1971 (MILLIONS OF DOLLARS) COMPUTER SERVICES INDUSTRY¹

YEAR	OFF-LINE	ON-LINE	TOTAL	VALUE OF COMPUTERS IN USE IN THE U.S.
1961	180		180	2800
1962	220		220	3800
1963	270		270	4900
1964	300		300	6400
1965	360		360	8100
1966	410	20	430	10800
1967	480	50	530	14000
1968	600	120	720	17900
1969	740	210	950	21900
1970	930	330	1260	25500
1971	1060	440	1500	28100

¹ ALL THE FIGURES IN THIS TABLE ARE ESTIMATES BY THE INTERNATIONAL DATA CORPORATION. THE VALUE OF COMPUTERS IN USE INCLUDES ALL COMPUTERS MADE BY AMERICAN-BASED COMPANIES IN PLACE IN THE UNITED STATES VALUED AT ORIGINAL PURCHASE PRICE AND IS FROM [28], PP. 6 AND 8, AND ITS EARLIER EDITIONS. THE COMPUTER SERVICES INDUSTRY TOTALS ARE ALSO FROM [28], P. 10, AND THE OFF-LINE AND ON-LINE ESTIMATES ARE FROM [29], SIXTH ANNUAL INDUSTRY STUDY, P. 4.

sectors emerged, both the growth pattern in the computer services industry and the relationship between the industry and the computer stock in use have changed. Using a piecewise linear regression on the Table 2 data, we found a significant break in the industry revenue trend between 1966 and 1967 and another in the industry revenue to computers in use relationship between 1967 and 1968.⁸ The growth

the Conference Board [14], in the broadest industry definition of all, would include broadcasting, publishing, and other information suppliers as well as the computer services in the information technology industry. The arguments for our choice are in the text.

⁶ Revenue estimates for this industry have to be considered quite unreliable. There is the common problem that firms operate in several sectors of the industry and outside of the industry, but to complicate matters further, the same is true of establishments. Each of the sources cited in Table 1 makes estimates of the industry, but all are forced to combine what they find in published financial reports for public corporations with what they can get from surveys and interviews and what they can work up from available figures on computers in use, rental estimates, and known revenue to rental ratios. In spite of this, the very general impressions of relative size and growth that emerge from the data seem quite acceptable to us and to people in the industry.

⁷ This is the high estimate for the computer services industry divided by the low estimate for electronic data processing expenditures from Table 1.

⁸ The technique and program we used are described in [43]. In our runs of the piecewise regression program for industry growth, the 1961-1966 cluster and the

rate changes from 18 percent for 1961 to 1966 to 30 percent for 1967 to 1971; the relationship changes so that a one-percent increase in the value of computers in use in the United States is associated with a 0.65-percent increase in industry revenue for 1961 to 1967 and a 1.65-percent change for 1968 to 1971. Its higher growth rate and increasing importance as an information source combined with the developing eclipse of the off-line conventional service bureaus by the on-line sector's⁹ direct policy attention to the industry.

But the basis for industry definition is the new services which are the source of its recent rapid growth and the rivalry with in-house computing which is the key to its future growth. The new services are a source of the low cross-elasticity of demand that now defines the industry. The computer services industry is moving into the old off-line services, and the computer manufacturers and others that serve in-house operations are moving into the new services. Approach and technology for each of the rivals are changing, and more and more customer-size groups are finding that in-house computing and the computer services industry are feasible alternatives for some or all of their information requirements. The promise of the new services and the possibility of a greatly changed structure of computer-based information sources are substantial arguments for policy interest in the computer services industry and its on-line sectors.

Whether new services will develop, whether in-house computing will largely be replaced by computer services, and whether the industry growth rate will remain above or fall below the growth rate in computer use cannot be predicted. But prediction is not our concern. Our concern is with the structure of the industry, that it not be altered unnecessarily in ways that would keep the industry from developing in the most publicly beneficial way.

■ Almarin Phillips uses the feedback from performance to structure to demonstrate a dynamic approach to the development of an industry.¹⁰ His equations and discussion suggest a framework for our analysis, and his emphasis on feedback leads to what might be considered the theme of our study: competition is an unnatural state for an industry. Competition, even if it produces performance judged more beneficial to the public than that of any alternative structure, is almost certain to leave the firms in the industry dissatisfied with the performance they relate to their particular objectives. Dissatisfaction encourages behavior that will cause structure and performance to change. Because movements toward competition are more difficult to achieve than movements away from it, change will be

3. Framework for analysis

1967-1971 cluster were the last two joined. The probabilities that the linear fit for either cluster was a matter of chance and that the two clusters were generated by the same system were all less than one percent. The same is true of the findings reported for the industry revenue to computer stock relationship.

⁹ International Data Corporation [27], p. A-11, Quantum Science Corporation [51], p. 39, and Arthur D. Little, Inc., [37], pp. 35, 44, all indicate that the on-line sectors will have more than half of industry revenues by 1977. Quantum Science, which has the most pessimistic view of off-line computer services industry opportunities, forecasts on-line revenues of \$1.8 billion and off-line revenues of \$750 million for 1977.

¹⁰ See Phillips, "Structure, Conduct and Performance—and Performance, Conduct and Structure," in [39], pp. 26-37, and [49].

toward a source of industry discipline more explicit, although not necessarily more effective, than the market.

Just how rapidly and how far an industry will move from competition will depend on the actions of the industry's machinery suppliers, its customers, and the units of government. The equipment suppliers may make decisions to influence industry structure, working against the structural decisions made by firms within the industry, but that influence and the motivation for it should be weak because of their interest in sales to other industries. Customers are the losers when structure moves away from competition, but the loss may not be apparent to them. Competition seems chaotic not only to firms in an industry but also to their customers. The promise of standardized products, controlled quality, established price structure, reduced firm turnover, and more responsible management associated with industry-initiated actions to change structure may be enough to win customer support for the changes. Some government units, the Anti-Trust Division, for example, may stand in the way of changes in industry structure, others, the Congress, and the regulatory commissions, may be the agents for changing structure. The chances are not great that any of these units can resist the pressure forcing an industry's structure away from competition, but they may improve if the structure and the pressures for change are made clear and the impact of alternative policy choices is at least indicated.

4. Three aspects of structure: cost, demand, and discipline

■ The structure of the computer services industry causes the industry to be competitive—even if not purely competitive.¹¹ The competitiveness involves the “independent striving for patronage by sellers in a market” and structural conditions that make that type of behavior difficult to avoid. There is also an implied judgment in our use of the term that “there is no clearly indicated change [from this type of competition] that can be effected through public policy measures that would result in greater social gains than social losses.”¹² In this section we discuss cost characteristics of the firms, the industry demand function, and the conditions that make it difficult to enforce anticompetitive measures. In the next section we point out the competition-derived pressures for change from within the industry and argue that there is social benefit in resisting them.

□ **Cost characteristics.** It is not really possible to find a typical firm for this industry. Services are differentiated in some way even among firms that seem very much alike, and the mix of services, computer processing technique, and method of delivering information varies within industry sectors as well as among them. Still, the firms share common cost characteristics. Here, we discuss these characteristics with particular reference to firms in remote job entry, time-sharing, and on-line dedicated applications.

Costs can be classified in various ways, but it is most useful to think of them as associated with development, capacity, and opera-

¹¹ See Chapter 2, [53], pp. 8–38, for a discussion of the distinction. Scherer does not tie Adam Smith and the concept of workable competition together but the association can be drawn from a reading of his opening and closing sections.

¹² Jesse Markham, “An Alternative Approach to the Concept of Workable Competition,” quoted in [53], p. 38.

tions. Development involves improving and creating software, and finding and educating customers. Capacity refers to the computer, other equipment, and the staff that must be on hand if services are to be available when customers call for them. And operations include supplies, technical staff, and sales representatives used when customers do call for services. Three different sources for estimating the distribution of these costs are represented in Tables 3, 4, and 5. The

TABLE 3
DISTRIBUTION OF EXPENDITURES ON ELECTRONIC DATA PROCESSING FOR INDUSTRIAL AND MANUFACTURING COMPANIES FROM DIEBOLD SURVEY, 1971¹ (PERCENT)

	DEVELOPMENT	CAPACITY	OPERATIONS	TOTAL
HARDWARE		35.4		35.4
SYSTEMS PERSONNEL	25.4			25.4
OPERATING PERSONNEL		14.3	14.3	28.6
SUPPLIES			6.8	6.8
EXTERNAL SERVICES			3.8	3.8
TOTAL	25.4	49.7	24.9	100.0
¹ THE DISTRIBUTION OF THE DIEBOLD ESTIMATES AMONG DEVELOPMENT, CAPACITY AND OPERATIONS IS OUR OWN. THE ALLOCATION OF ALL SYSTEMS PERSONNEL TO DEVELOPMENT MAY EXAGGERATE OUTLAYS A LITTLE FOR THIS ACTIVITY, BUT IT DOES GAIN SUPPORT FROM DIEBOLD'S SURVEY RESULTS FOR THE ALLOCATION OF EDP RESOURCES TO NEW SYSTEMS, ONGOING APPLICATIONS AND CONVERSION. THE ALLOCATION IS 30.4 PERCENT NEW SYSTEMS DEVELOPMENT, 59 PERCENT ONGOING APPLICATIONS, AND 10.6 PERCENT CONVERSION. SPLITTING OPERATING PERSONNEL OUTLAYS HALF AND HALF BETWEEN CAPACITY AND OPERATIONS IS ARBITRARY, BUT WE FEEL THAT ANY UNDER-STATEMENT IN THIS WILL BE IN CAPACITY OUTLAYS				
SOURCE: [17], P. 15.				

variations among the sources and among the 16 remote sellers reported in the final table are associated with real differences in activity as well as with estimation problems and should be noted. But it seems reasonable to summarize these sources by saying that costs for firms providing computer services tend to be distributed 20 percent for development, 50 percent for capacity, and 30 percent for operations.

This distribution makes it worthwhile to pay some attention to the distinction between cost or expense and expenditure. Cost or expense associates resource use with a revenue producing activity, while expenditure simply notes that one resource is transformed into another that will, in some later activity, have its value exhausted. The purchase of plant is an obvious expenditure, and depreciation, an attempt to associate the expense or cost with the appropriate activity. Because development and capacity costs are so important and because the cycle from initiation to full operation may be two, three, or four years for a computer services firm, expenditure rather than cost or expense is likely to appear in accounting statements. In its early years a firm will plan to do software development, promotion, and education that will produce activity and reduce operating cost in later years. It may also plan for more capacity than it seems to need in the early years, because a single effort at establishing and integrating capacity may produce savings. These actions involve expenditure that should be capitalized, not expensed. The securities market will do just that if it expects the plans to be realized. Accounting statements will not, however, and they have to be viewed accordingly.

Expenditure depends on plans; it also depends on realized activity, but to a much lesser extent. The type of activity planned will influence expenditure in two distinct ways. Higher quality service,

TABLE 4

DISTRIBUTION OF EXPENDITURES BY COMPUTER SERVICES FIRMS FROM ADAPSO SURVEY, 1971¹ (PERCENT)

	DEVELOPMENT	CAPACITY	OPERATIONS	TOTAL
EQUIPMENT		30		30
LABOR ²				(42)
EXECUTIVE, MANAGEMENT	2	2	2	6
SALES	3		5	8
PROGRAMMERS	2		2	4
SYSTEMS ANALYSTS	6			6
OPERATORS		4		4
KEYPUNCH	1		3	4
CLERICAL	1		4	5
OTHER	1	2	2	5
SUPPLIES			4	4
EXTERNAL SERVICES			2	2
OVERHEAD		7		7
ADVERTISING	3		3	6
OTHER	3	3	3	9
TOTAL	22	48	30	100

¹ THE DISTRIBUTION OF THE ADAPSO ESTIMATES AMONG DEVELOPMENT, CAPACITY AND OPERATIONS IS OUR OWN. THERE IS PROBABLY LITTLE THAT WE HAVE PLACED IN THE CAPACITY COLUMN THAT COULD BE TAKEN AWAY. THERE COULD BE A SHIFT OF PERHAPS AS MUCH AS SIX PERCENT BETWEEN DEVELOPMENT AND OPERATIONS

² THE ESTIMATES FOR THE LABOR SUB-GROUPS WERE MADE BY TAKING THE PERCENT OF EMPLOYEES IN EACH GROUP [29], p. 21, WEIGHTING THE PERCENT OF SALARIES OF EACH GROUP [29], p. 22, AND DISTRIBUTING THE LABOR OUTLAYS ON THIS BASIS. THE FIGURES USED ARE AS FOLLOWS:

	PERCENT OF EMPLOYEES	SALARY PER WEEK
EXECUTIVE, MANAGEMENT	15	
SALES	7	\$273
PROGRAMMERS	13	217
SYSTEMS ANALYSTS	6	275
OPERATORS	13	146
KEYPUNCH	23	107
CLERICAL	19	110
OTHER	4	97

IT WAS ASSUMED THAT THE SIX PERCENT EXECUTIVE AND NINE PERCENT MANAGERIAL-SUPERVISORY WHO MADE UP THE EXECUTIVE MANAGEMENT GROUP RECEIVED \$500 AND \$400 A WEEK RESPECTIVELY.

SOURCE [29], SIXTH ANNUAL INDUSTRY STUDY, 1972.

more versatile, responsible, or reliable service, will require a larger expenditure.¹³ A wider mix of services, combining several industry sectors, software development, consulting, and facilities management, for example, will result in a lower expenditure for each service.¹⁴ The distribution of activity planned will also influence expenditure. Distributing the activity among more customers or more locations will increase expenditure; distributing it more evenly over the hours of the year will decrease expenditure. Finally, the scale of activity planned will have an effect. For small firms expansion offers savings, but economies of scale are exhausted before any sizable share of the market is involved. Expenditure becomes cost, and plans consider all possible changes and so represent the long-run. Cost character-

¹³ Price variations among firms with the same computer systems in the time-sharing sector are a reflection of this. The comparisons we have made are discussed below.

¹⁴ Casual use of the "survivor" approach [60] supports this view. Not only are most firms listed in the September 1972 Time Sharing Directory [62] involved in several industry sectors, but firms that appear in this directory and in earlier ones have tended to add activities within and outside the industry.

TABLE 5

DISTRIBUTION OF EXPENDITURES FOR 16 REMOTE SELLERS OF COMPUTER SERVICES (PERCENT)

FIRM NUMBER ¹	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
FISCAL YEAR ENDING	Dec71	Sep71	Jun71	Jun71	May71	Jul71	Dec71	Feb72	Apr71	Nov71	Dec71	Jun71	Oct71	May71	Dec71	Feb72
REVENUE (\$000)	546	2461	4149	5280	2635	6096	2735	11109	2466	1635	5401	4085	333	231	12520	3000
DEVELOPMENT ²	10	9		5	13	22	8	15	6	4	15	10	11	2	9	32
SOFTWARE	10	9		5		22	8	5	6	4		10				11
MARKETING																21
SALES LABOR											5			2		
SYSTEMS LABOR								10			10					
DEVELOPMENT					13								11		9	
CAPACITY ²	55	59	69	63	58	49	52	56	58	57	49	56	55	63	64	40
COMPUTER EQUIP. ³		30**	23*	31	18*	23*	7*	31	22*	19**	21**	22	21	20	33	18**
OTHER EQUIP.			6		11		4	15			18				4	
OPERATIONS	40	9	20			19	23	10	9	28	10	16		10	7	6
SELL, GEN. ADMIN	15	20	20	32	22	7	18		27	10		18	23	33	20	16
DEVELOPMENT					13								11			
OPERATIONS ²	35	32	31	32	29	29	40	29	36	39	36	34	34	35	27	28
OPERATIONS	20	10	10			19	22	11	9	28		16			6	6
SALES LABOR											5			2		
SELL, GEN. ADMIN.	15	20	21	32	23	7	18		27	11		18	23	33	21	
MARKETING																22
OTHER		2			6	3		18			31		11			
OUTLAYS (\$000)	602	4306	5051	6282	2690	6758	2384	10323	1974	1573	4409	5575	368	484	11960	2360
GAIN (LOSS)	(9)	(43)	(18)	(16)	(2)	(10)	15	8	25	4	23	(26)	(6)	(52)	5	27
(% OUTLAYS) ⁴																

¹ THE FIRMS, WITH THE DATE OF FIRST OPERATION, ARE

- | | |
|---|--|
| 1. ACADEMY COMPUTING CORPORATION; FEB., '68 | 10. PROGRAMS AND ANALYSIS, INC., JAN., '69 |
| 2. APPLIED LOGIC CORPORATION; JAN., '66 | 11. RAPIDATA, INC.; FEB., '68 |
| 3. COMPUTER COMPLEX, INC.; JULY, '67 | 12. SYSTEMS DIMENSIONS, LTD., JAN., '69 |
| 4. COM-SHARE, INC.; JUNE, '66 | 13. TIME SHARE CORPORATION; JULY, '69 |
| 5. INTERNATIONAL TIME SHARING CORPORATION; MAY, '68 | 14. TIME SHARING RESOURCES; JAN., '70 |
| 6. KEYDATA CORPORATION; '65 | 15. TYMSHARE, INC.; NOV., '66 |
| 7. METRIDATA COMPUTING, INC.; JAN., '69 | 16. NAME WITHHELD, NONPUBLIC; SEPT., '69 |
| 8. NATIONAL CSS, INC.; DEC., '68 | |
| 9. ON-LINE SYSTEMS, INC.; JAN., '68 | |

² ALL OUTLAYS ARE EXPRESSED AS A PERCENT OF TOTAL OUTLAYS. TOTAL OUTLAYS WERE CALCULATED BY TAKING REPORTED EXPENSE, EXCLUDING INTEREST, TAXES, OR EXTRAORDINARY CHARGES, AND ADDING TO IT THE DIFFERENCE BETWEEN WHAT WAS SPENT FOR SOFTWARE DEVELOPMENT AND WHAT WAS EXPENSED. DEPRECIATION WAS TAKEN AS AN OUTLAY FOR FIRMS THAT OWNED THEIR COMPUTERS OR OTHER PROPERTY. THE EXPENDITURE CATEGORIES ARE APPROXIMATELY THE SAME AS THOSE USED IN A FIRM'S ANNUAL REPORT. THE ABSENCE OF AN ENTRY IN A PARTICULAR CATEGORY SIMPLY MEANS THAT THE FIRM DID NOT USE THAT CLASSIFICATION. THE DISTRIBUTION AMONG DEVELOPMENT, CAPACITY AND OPERATIONS IS OUR OWN. WHEN CONFRONTED WITH EXPENSE CATEGORIES SUCH AS OPERATIONS, SELLING, GENERAL OR ADMINISTRATIVE AND NO OTHER GUIDANCE, WE SIMPLY DIVIDED THEM HALF AND HALF TO CAPACITY AND OPERATIONS. THIS PROBABLY CAUSES DEVELOPMENT TO BE SOMEWHAT UNDERSTATED.

³ BOTH OUR OWN DIRECT MAIL SURVEY OF FIRMS AND [62] REPORT ON THE COMPUTER EQUIPMENT USED BY FIRMS. AVERAGE MONTHLY RENTALS FOR GENERAL-PURPOSE COMPUTER SYSTEMS IN USE ARE FOUND IN [28]. USING THESE SOURCES TOGETHER, ESTIMATES CAN BE MADE OF COMPUTER EQUIPMENT OUTLAYS IN THE TABLE.** AFTER THE COMPUTER EQUIPMENT FIGURE MEANS THE COMPUTER EQUIPMENT RENTAL WAS NOTED IN THE ANNUAL REPORT AND WAS CONSISTENT WITH THE ESTIMATE FROM THESE SOURCES. A * AFTER THE FIGURE MEANS THE COMPUTER EQUIPMENT RENTAL WAS NOT IN THE ANNUAL REPORT, AND WE USED THE ESTIMATE FROM THESE SOURCES TO SPLIT THIS OUTLAY OUT OF FIGURES FOR OPERATIONS. IF THERE IS NOTHING AFTER THE FIGURE IT IS DIRECTLY FROM OUR ESTIMATE FROM OTHER SOURCES. THE DIFFERENCE MAY BE DUE TO SPECIAL LEASE NEGOTIATIONS, DEPRECIATION ON A LONGER THAN LEASE BASIS, OR TO THE FACT THAT THE SYSTEM IN USE HAS A CONFIGURATION THAT DIFFERS FROM THE AVERAGE. IF THERE IS NO FIGURE AT ALL, THEN NEITHER THE ANNUAL REPORT NOR THE OTHER SOURCES PROVIDED A REASONABLE ESTIMATE.

⁴ GAIN OR (LOSS) IS CALCULATED BY SUBTRACTING EXPENDITURE FROM REVENUES AND DIVIDING THE RESULT BY EXPENDITURES.

SOURCE: PUBLISHED ANNUAL REPORTS; DIRECT SURVEY OF FIRMS

istics, therefore, emerge from this brief description of expenditure determinants. The most important of these for our purposes are very limited economies of scale, long-run incremental cost that varies little with off-peak business, and, in the short-run, very high fixed cost.

The case seems clear for high fixed cost in the short-run where, by definition, plans and capacity remain unchanged. We have already noted that 70 percent of expenditure is associated with development or capacity. These expenditures for rent or depreciation of the computer system and other equipment, the system operating staff, software development, market promotion, and administration may change with plans for the future but should not change if activity varies from expected levels for a quarter or even a year. Among operating expenditures those associated with commissions, sales representatives, technical staff, supplies, and communications will change, but even in this group the presence of supervisors and superstructure makes the response to a temporary change in activity less than proportional. Cost involves more than current expenditure, however, and there is reason to expect short-run marginal cost to rise sharply as activity in peak hours approaches capacity. The people involved in software development and market promotion have some skills in common with the technical staff and sales representatives who respond directly to customers. In addition, their activities require the computer as input, and they compete with customers for machine time during peak hours. As the level of activity rises, these programmers and market developers will be shifted to operations, and their contribution to future business and future systems efficiency will be lost. Even though there will be a balance struck between this opportunity cost and the cost of adding technical staff and sales representatives temporarily, short-run marginal cost will rise as shifting increases. The rise will be accentuated by the loss in current and future sales which results when customers, vying for computer time during peak hours, find themselves forced to wait for service or getting less service for every computer hour they pay for. Marginal cost that is low for most of the output range and that rises sharply only when capacity is approached is, of course, just another view of high fixed cost in the short-run.

To support or estimate these short-run relationships empirically, beyond drawing attention to the distribution of expenditure, is difficult. The short-run is a concept, not an observable period. In the periods that can be observed, expenditure rather than cost is reported, and the expenditure is associated with plans that run several periods and that can change from period to period. A further complication is the absence of any reasonable measure of the quantity of information produced and, for commercial firms, any available statistics on possible output proxies such as the hours of central processing unit time used by customers. The barriers to direct estimation of short-run cost relationships do not rule out consideration of the type of data that are available, and these data seem quite consistent with our description.

Revenue and expenditure information for fifteen remote sellers of computer services is presented in Table 6. Expenditure is calculated by taking operating expense reported by a firm, deducting interest changes, and adding in the difference between outlay for and amorti-

TABLE 6

REVENUE AND EXPENDITURE FOR 15 REMOTE SELLERS OF COMPUTER SERVICES

FIRM	YEAR	REVENUE (\$000)	EXPENDITURE (\$000)	$\frac{\Delta \text{EXP}}{\Delta \text{REV}}$	$\frac{\Delta \text{REV}}{\text{REV-1}}$	$\frac{\text{REV-EXP}}{\text{EXP}}$
1. ACADEMY COMPUTING CORP	1970	350	457			-0.23
	1971	546	602	0.74	0.56	-0.08
2. APPLIED LOGIC CORP.	1967	533	441		1.21	0.25
	1968	1176	1054	0.95	1.63	0.12
	1969	3090	3630	1.35	-0.19	-0.15
	1970	2508	8504	*		-0.71
	1971	2461	4306	*	-0.02	-0.43
3. COMPUTER COMPLEX, INC.	1968	239	1140			-0.79
	1969	2007	4305	1.79	7.40	-0.53
	1970	4113	5574	0.60	1.05	-0.26
	1971	4149	5057	$\frac{-517}{36}$	0.01	-0.18
4. COM-SHARE, INC	1968	1117	3402			-0.67
	1969	3966	6942	1.24	2.55	-0.44
	1970	4879	7498	0.61	0.23	-0.54
	1971	5280	6282	*	0.08	-0.16
5. INTERNATIONAL TIME SHARING CORP.	1969	458	1721			-0.73
	1970	1835	2538	0.59	3.01	-0.28
	1971	2635	2690	0.19	0.44	-0.02
6. KEYDATA CORP.	1968	2030	3774			-0.46
	1969	2455	4204	0.21	1.01	-0.42
	1970	4359	5728	0.78	0.80	-0.24
	1971	6096	6758	0.40	0.59	-0.10
7. METRIDATA COMPUTING, INC.	1969	378	898			-0.58
	1970	1584	1622	0.60	3.19	-0.02
	1971	2735	2384	0.66	0.73	-0.13
8. NATIONAL CSS, INC.	1968	135	439			-0.69
	1969	2587	3484	1.24	18.16	-0.26
	1970	7790	10040	1.26	2.01	-0.22
	1971	11109	10323	0.09	0.43	-0.08
9. ON-LINE SYSTEMS, INC.	1967	33	319			-0.90
	1968	474	651	0.75	13.36	0.27
	1969	1588	1385	0.66	2.35	0.15
	1970	2466	1974	0.80	0.55	0.25
	1971	3525	—	0.96	0.49	0.18
10. PROGRAMS AND ANALYSIS, INC.	1969	939	1032			0.09
	1970	1398	1379	0.76	0.49	0.01
	1971	1635	1573	0.82	0.17	0.14
11. RAPIDATA	1967	0	90			-1.00
	1968	470	784	1.48	470	0.40
					0	
	1969	2079	2434	1.03	3.42	0.15
	1970	4376	4520	0.91	1.10	0.03
	1971	5401	4409	*	0.23	0.23
12. SYSTEMS DIMENSIONS, LTD.	1969	0	9			-1.00
	1970	1330	4031	3.03	1330	0.67
					0	
	1971	4805	5575	0.56	2.61	0.26

TABLE 6 (CONTINUED)

FIRMS	YEAR	REVENUE (\$000)	EXPENDITURE (\$000)	$\frac{\Delta EXP^1}{\Delta REV}$	$\frac{\Delta REV}{REV-1}$	REV-EXP EXP
13. TIME SHARE CORP.	1966	411	410			0
	1967	863	902	1.09	1.10	0.04
	1968	1158	1434	1.80	0.34	0.19
	1969	3241	3892	1.18	1.80	0.17
	1970	392	1081	*	0.88	0.62
	1971	333	368	*	0.15	0.06
14. TIME SHARING RESOURCES, INC.	1969	108	213			-0.49
	1970	231	484	2.20	1.14	-0.52
	1971			0.48	1.17	-0.28
15. TYMSHARE, INC.	1966	112	351			0.68
	1967	1002	1304	1.07	7.95	0.23
	1968	2618	3458	1.33	1.60	0.24
	1969	6372	6962	0.93	1.43	0.08
	1970	10233	9936	0.77	0.61	0.03
	1971	12520	11960	0.88	0.22	0.05

¹ IN ROWS MARKED * REVENUE AND EXPENDITURE MOVE IN OPPOSITE DIRECTIONS OR ARE BOTH NEGATIVE. THE REVENUES AND EXPENSES, IN MILLIONS OF DOLLARS ARE AS FOLLOWS: APPLIED LOGIC CORP., 1970 REVENUE +4874, EXPENDITURE -582; APPLIED LOGIC CORP., 1971, -4198, -47; COMPUTER COMPLEX, INC., 1971, -516, +36; COM SHARE, INC., 1971, -1216, +401; RAPIDATA, 1971, -111, 1025; TIME SHARE CORP., 1970, -2811, -2849; TIME SHARE CORP., 1971, -713, -59.

SOURCE: PUBLISHED ANNUAL REPORTS AND PROSPECTUS.

zation of software and development. This provides an approximation of what a firm would have spent if it leased its equipment and rented its facilities, a combination of actions typical of these firms.¹⁵

The last column of the table shows that, for all but two of these firms, the operating deficit is substantial in the early years of operation, then gradually declines and tends to disappear as time goes on. This is consistent with a plan of several years' duration where early expenditure is required to make later revenue possible. Plans change from year to year, however, and there is some evidence in the table that the changes are made quickly and that they significantly influence spending. There are 38 cases where expenditure and revenue both increase. In fourteen cases, $\Delta EXP/\Delta REV$ is greater than one. Thirteen of these involve revenue increases of more than 100 percent; eleven occur through 1969. The 24 cases with $\Delta EXP/\Delta REV$ less than one include only nine where revenue growth is more than 100 percent and five from the years through 1969. In the earlier years when growth was more rapid, experience was better than expectation, and plans were promptly revised with the observed impact on incremental spending. In 1971 the situation was reversed. Experience was below expectation, and five of the fifteen firms changed plans and cut expenditure, three with revenue increasing and two by more than \$10

¹⁵ Omitting explicit interest from operating expenses for a firm that owns its computer while making no estimate of the interest portion of a rent payment for a firm with a long-term computer lease creates some discrepancy in a cross-sectional comparison. The problem is a minor one, however, not only because the charges involved are relatively small and the data full of imperfections to begin with, but also because the leases are often cancellable with a small penalty payment and so may not be properly looked upon as long-term financial arrangements.

for each \$1 of revenue drop. This emphasizes again that the observed changes in expenditure from period to period relate to steps in a single plan or changes from one plan to another, not to short-run shifts in output.

Expenditure dominated by plans rather than actual output is what we should observe if fixed cost is high in the short-run. If, instead, most cost were variable, then a change in output and revenue would have a closely matched change in cost. The observed distribution in $\Delta EXP/\Delta REV$ would be narrow and, in the extreme case of no fixed cost, would be a good estimate of marginal cost. The very wide range of the distribution in the $\Delta EXP/\Delta REV$ column is consistent with the claim that this is an industry where the greatest share of cost is associated with standing ready to serve and is fixed in the short-run.

Somewhat more direct evidence of this is available from observations of Dartmouth College's Computation Center where we do have information on terminal hours and central processing unit seconds available and put to use.¹⁶ This information is used to estimate activity and capacity. Expenditure is estimated in the same manner as for the commercial firms. The year-to-year percentage changes are as follows:

	<u>1968-1969</u>	<u>1969-1970</u>	<u>1970-1971</u>	<u>1971-1972</u>
Activity	+15%	+3%	+54%	+17%
Capacity	0	+8%	+41%	0
Expenditure	+ 4%	+8%	+30%	- 7%.

These data, sketchy as they are, offer a further indication that cost varies very little with output in the short-run.

The long-run relationships have already been described and most of them do not require further discussion. Three that do and that have something of a connection involve diversification, peak loading, and scale.

The claim that the cost of a particular service is lower when the service is produced by a diversified firm had some support in an earlier footnote but deserves brief comment.¹⁷ This relationship is strongest when scale is small. Members of the software development staff, particularly the most expert systems programmers, are not utilized fully at all times in the computer service. They can carry on some software work for outsiders as well. The same is true of analysts who provide decision-making routines and consult with customers. They can also undertake independent consulting. Even the computer, the center of a system to provide remote services, may have spare hours to be devoted to off-line batch processing. To these savings from indivisibilities can be added some educational feedback that may turn into new systems or better algorithms from an improved staff and some savings associated with applying a particular bit of intellectual output more widely. The evidence that savings of this type exist and may provide an alternative to expanding scale seems to be suggested by the firms that survive in the industry. Results from our survey of remote sellers, shown in Table 7, suggest that very few firms paying \$100,000 or less in monthly computer system

¹⁶ The details of this study are contained in [8].
¹⁷ See footnote 14.

TABLE 7

DIVERSIFICATION OF REMOTE SELLERS INTO OFF-LINE BATCH PROCESSING, SOFTWARE DEVELOPMENT, FACILITIES MANAGEMENT AND OTHER SERVICES¹

MONTHLY RENT FOR COMPUTER SYSTEM (\$000)	REMOTE ONLY	NUMBER OF REMOTE SELLERS OFFERING			TOTAL
		OFF-LINE BATCH	OTHER ²	OFF-LINE BATCH AND OTHER	
1 TO 40	2	6	5	12	25
41 TO 100	1	2	3	9	15
101 TO 300	3	3	2	11	19
301 AND UP	1	0	1	3	5
ALL	7	11	11	35	64

¹THE DATA ARE FROM 64 USABLE QUESTIONNAIRES RETURNED IN OUR MAIL SURVEYS OF 99 FIRMS IN THE REMOTE JOB ENTRY, TIME-SHARING AND ON-LINE DEDICATED APPLICATIONS SECTORS OF THE COMPUTER SERVICES INDUSTRY. THERE WERE TWO SURVEYS. ONE IN 1971 AND ONE IN 1972. THE 33 FIRMS THAT REPORTED THE DISTRIBUTION OF THEIR REVENUES INDICATED A RANGE OF 5 PERCENT TO 93 PERCENT OF REVENUES FROM OFF-LINE BATCH AND 3 PERCENT TO 60 PERCENT OF REVENUES FROM OTHER SERVICES.

²OTHER AS USED HERE INCLUDES ONLY SOFTWARE DEVELOPMENT, CONSULTING, FACILITIES MANAGEMENT AND EDUCATION

rentals survive when their remote services are not mixed with off-line batch, software development, consulting, facilities management, or education.

There is a very great difference in the incremental cost of peak and off-peak business. If one of fifty customers to be provided service promptly and reliably between 3 and 4 P.M. on a particular day doubles his requirements, and if fifty represents the computer system's peak load for an hour, then the only way to satisfy the customer is to expand the system. This will involve adding to development, capacity, and operating cost. Just how much is a question of scale economies which we shall come to shortly. If that customer were to double his requirements but take the additional service between 9 and 10 P.M. when only five other customers were using the service, then only a small portion of operating cost would change. The importance of the computer and the staff to program, run, and maintain it makes this peak load relationship apparent.

But communication cost has a similar influence involving not only time but place. To do business in a city away from its computer, a firm that sells remote services requires a phone line to the city, termination arrangements at each end of the line, and equipment that packs the line efficiently with data.¹⁸ The line charge is fixed by distance. And one private, voice-grade line can be packed to handle simultaneously at least 48 computer system users transmitting and receiving at conventional teletype speed. The termination and equipment charges vary with the number of customers the firm stands ready to serve. Activity added off-peak from any city that is on line

¹⁸ This description ignores the possibility of using WATS, wide area telephone service. The omission is significant because WATS is the most cost effective way to offer remote service to customers that are widely dispersed. Although it is the best way to handle 50 customers each located in a different city, it will result in much higher communication costs than that of handling 50 customers in one city by direct private line.

has no effect on communication cost. Activity added at a city that is not fully utilizing a line, even if at the peak hour, has an incremental cost that excludes line charges and that will also be below average unit cost for termination and equipment. Dispersing business through time and concentrating it geographically mean lower communication cost.

The relative importance of communication cost affects the extent of the market and the influence of Federal Communications Commission decisions. To serve simultaneously twelve users 300 miles away over a voice-grade private line sliced so that each user will transmit on his own frequency involves monthly charges of:¹⁹

Equipment Depreciation	\$430
Equipment Maintenance	300
Termination, Business Phones, Data Sets	477
Private Voice-Grade Line (300 miles)	581
TOTAL	\$1788.

If the twelve channels are in use six hours a day, twenty days a month, which might be considered normal, this would be \$1.24 for each hour a user is on-line with the computer. Firms in the time-sharing sector of the industry using a Honeywell G400 series computer now charge \$11 an hour to hook up and, perhaps, another \$5 each hour for central processing unit and storage.²⁰ The communication cost for the computer services firm in this case would be about eight percent of the hourly charge to the customer. If the customers were 3000 miles away, the line would cost an additional \$2350, the hourly communication cost would be \$2.89, and the percentage of hourly charge would be about 18 percent. The same 3000 mile line could be split among 48 users. Equipment costs would not quadruple in going to 48 users because another method of packing the line would be necessary and advantageous.²¹ However, if we assume that equipment and termination costs simply quadruple and if use per channel remains the same, then the hourly communication cost for users 3000 miles away drops to \$1.35, less than 10 percent of the hourly charge.

The 8 to 18 percent communication's share of cost suggested by our examples is probably a good approximation for most firms in the time-sharing sector.²² Rapidata, the only firm that provides data that permit separation, devoted 9 percent of expenditure to communication in 1971.²³ There are firms in this sector, Tymshare is an example, that will offer more and better communication in their mix of services; a larger share of their expenditure will be spent on communication.²⁴ In other sectors the share will also vary. Data bank reference systems will use more communication and less computer time because retrieval and screening dominate their activity. On-line dedicated applications and remote job entry will use less communication

¹⁹ The figures are based on the actual experience of the Dartmouth Computer Center. A full discussion of them is contained in [8].

²⁰ The figures are for Rapidata and are from [62].

²¹ For a discussion of this operation see [41], Ch. 15.

²² A recent Diebold Research Program study [18] offers some corroboration.

²³ Data provided by Rapidata.

²⁴ A discussion of remote computing networks appears in [63] and [41].

and more computer time because they are more likely to keep customers off-line while doing calculations. Communication is the cost of transporting information. In all of these sectors it limits a firm's market, not in distance, but to locations where customers are concentrated, and it offers some incentive for larger scale.

In spite of the fact that communication is an incentive for larger scale rather than a barrier, economies of scale are not important in the industry. This is not because larger computers offer no advantage. Nor is it because greater volume fails to conserve start-up efforts, limit divisibility problems, and permit learning. If information of a narrowly specified type and form is produced, unit cost will fall as volume is increased by replicating a system making use of a large computer. The difficulty is that one particular type and form of information is not quite right for the different needs within and among organizations. The different needs call for different computer systems. Of course, one firm could operate a great many systems and could direct each user and each use to the most appropriate one. But the mechanical and managerial techniques for doing this are at a stage where the cost of direction rises rapidly as the burden of more and more choices is assumed. As a result, many choices have to be passed back to users. Because they are, any economic advantage of combining many different types of computer systems within a single firm disappears. The result is that once a computer services firm expands beyond customers for a particular type of information in a particular package, its size offers no advantage.²⁵

Evidence to support this comes, unexpectedly, from a study of computer system size. Grosch's law states that a one-percent increase in computer system rent is associated with a two-percent increase in system power. The many studies that support this²⁶ compare rent and power of various computer systems by assuming that computer power, measured by multiplication speed, throughput, memory size, or other characteristics, represents output. However, because the product is information, not the mechanics of information production,²⁷ it may be just as well to assume as Chow does²⁸ that rent is the measure of a system's capacity to produce information. This assumes away economies of computer system scale in any study that

²⁵ This paragraph argues that the long-run average total unit cost curve in computer services is *L*-shaped. This is a common finding. See [33]. Ours is not a common case for the *L* shape, however. It is really based on recognizing that information, not computer power, is the product. A more homey example of the same point could be made with regard to restaurants. If Mexican food of a particular type and form would satisfy every restaurant patron in New York City, we would expect a single chain of restaurants to offer each dinner at a lower cost than a set of individually owned restaurants. Restaurant patrons are not satisfied with such uniform fare. As a result, there are many restaurants, each slightly different from others, operating in the New York market. Perhaps in some strict sense, for a badly defined product, there are continuing economies of scale. In a much more reasonable sense there are no continuing economies, because the real market for restaurant or computer services is not open to their realization.

²⁶ See [56], pp. 314-322.

²⁷ In [23], Haldi and Whitcomb discuss the cost of basic industrial equipment as one source of scale economies. Their findings on equipment prices and scale are similar to the findings for computers. The types of equipment they consider have a scale more easily measured in terms of output, but they also may have some problems of distinguishing the product from its method of production.

²⁸ See [13].

compares rent and power across systems. The question of scale economies for computer systems, therefore, cannot be resolved by this type of comparison. It requires considering the choices of computer users. Lee Selwyn has done this, and his data support our conclusions on scale.²⁹

Selwyn looked at computer use across industries and fit the following equation cross-sectionally to 1966 data for 119 four-digit industries:

$$R/N = b_0 S^{b_1} C^{b_2} P^{b_3},$$

where R/N is average rent of a computer in an industry, S is total value of shipments, C is the four-firm concentration ratio, and P is the number of individual establishments operated by the four largest firms. He found $b_1 = +0.158$, $b_2 = +0.261$, and $b_3 = -0.141$, all with significance at the one-percent level. Assuming that information requirements were proportional to sales, Selwyn looked at the coefficient of b_1 and concluded that a one-percent increase in information required only a 0.158-percent increase in computer system size, indicating very substantial scale economies. He neglected the fact that the number of systems as well as the average size of a system changes. Using the number of systems rather than average rent as a dependent variable with Selwyn's data and independent variables, then $b_1 = +0.929$, $b_2 = +0.192$, and $b_3 = -0.251$, all significant at the one-percent level. Combining results of the two regressions, we have an indication that, as information requirements increase, they are met in part by increasing the size of computer systems and to a much greater extent by increasing the number of systems.³⁰

It might be added that the type as well as the number of computer systems increases. The 1966 data indicate that the aircraft industry, with 67 percent of output produced in nine establishments, used 455 computer systems. McDonnell-Douglas in 1971 reported having 109 computers,³¹ and Boeing Computer Services, the computer arm of Boeing Aircraft, listed 88 computers of 24 different types in the same year.³² All this suggests that the most efficient response to increased information requirements is not simply a larger computer or a replicated system. It does not demonstrate that there are no scale advantages in running many different systems. The best evidence of this is in the performance of firms in the computer services industry.

²⁹ See [54].

³⁰ An industry cross section may not be an ideal basis for this point, although the computers in an industry may provide information that goes beyond the firms or establishments in which they are located. We also used Selwyn's data to derive firm and establishment size for the big four in each industry and ran simple regressions across industries relating average rent and number of computers to big four firm or establishment size. The regression coefficients with firm size as the independent variable were +0.163 for average rent and +0.510 for number of computer systems. With establishment size the coefficients were +0.189 for average rent and +0.395 for number of computer systems. The data in all these cases are not quite what are wanted but the picture is consistently the one we describe in the text.

³¹ Reported in a speech, "The Planning for and Development of the McDonnell-Douglas Automation Company," by A. Joseph Quackenbush, Executive Vice President of McDonnell-Douglas Automation Company, at the Diebold Research Program Twenty-Fifth Preliminary Meeting, San Francisco, California, March, 30, 1971.

³² Response to our direct inquiry.

In July 1971 *Time-Sharing Today* reported 28 profitable companies selling remote computer services and asked, "What type of company is profitable?" It noted³³:

Profitable companies seem to come in all sizes. The revenues of our 28 companies range from \$100,000 to \$20,000,000 per year:

Remote-Computing Revenues	Companies
\$ 100,000-\$ 1,000,000	10
1,000,000- 5,000,000	12
5,000,000- 10,000,000	3
10,000,000- UP	3.

Performance information for 14 of the 16 firms we introduced earlier is shown in Table 8.³⁴ The firms are ranked by outlay which is used

TABLE 8

PERFORMANCE RATIOS AND SCALE (MILLIONS OF DOLLARS)

FIRM NAME	TOTAL OUTLAYS FISCAL YEAR 1971	REVENUE FISCAL YEAR 1971	RATIO OF REVENUE TO EXPEND.	EQUITY INVESTMENT ¹ FISCAL YEAR 1971	MARKET VALUE ² DECEMBER 1971	RATIO OF MARKET VALUE TO EQUITY INVESTMENT	MONTHS SINCE START OF OPERATIONS ³
TYMSHARE, INC.	\$11.960	\$12.520	1.05	\$7.01	\$24 01	3.42	62
NATIONAL CSS, INC.	10 323	11.109	1.08	5.03	8 61	1.71	39
KEYDATA CORPORATION	6.758	6.096	0.90	7.95	15.58	1.95	72
COM-SHARE, INC.	6.282	5.280	0.84	8 78	4.32	0.49	61
COMPUTER COMPLEX, INC.	5.057	4 149	0.81	7.00	0.35	0.05	49
RAPIDATA, INC.	4.409	5 401	1.22	1.22	8.84	7.25	47
APPLIED LOGIC CORP.	4 306	2 461	0.57	6 88	0.54	0.08	69
INTERNATIONAL TIME SHARING CORP.	2.690	2.635	0.98	5.10	8.91	1.75	37
METRIDATA COMPUTING, INC.	2 384	2.735	1.15	1 87	3.70	1.98	36
ON-LINE SYSTEMS, INC.	1.974	2.466	1.25	1.75	4.39	2 51	40
PROGRAMS & ANALYSIS, INC.	1 635	1.573	1.04	0 15	2.75	18.33	35
ACADEMY COMPUTING CORP.	0.546	0 602	0.91	2 70	1.55	0.57	43
TIME SHARING RESOURCES	0.231	0.484	0.48	0.75	1.55	2.07	17
TIME SHARE CORPORATION	0 333	0 368	0.90	2 17	0.82	0.38	28

¹THIS IS THE INVESTMENT MADE BY EQUITY INVESTORS TO THE END OF THE 1971 FISCAL YEAR.

²THIS IS BASED ON A DECEMBER 1971, STOCK PRICE QUOTATION FROM THE BANK AND QUOTATION RECORD OR NATIONAL MONTHLY STOCK SUMMARIES. THE STOCK PRICE IS MULTIPLIED BY SHARES OUTSTANDING AT THE END OF THE FISCAL YEAR TO GET MARKET VALUE. FISCAL YEARS VARY. ALL FALL WITHIN THE MONTHS APRIL 1971 TO MARCH 1972 WITH THE EXCEPTION OF TIME-SHARING RESOURCES, MAY 1972.

³NUMBER OF MONTHS FROM THE START OF OPERATIONS TO THE CLOSE OF THE 1971 FISCAL YEAR

as a rough measure of size.³⁵ The two performance measures provided are the revenue to expenditure ratio and the market to book value ratio. The first is some improvement on reported profits but has problems we have already discussed. The second is an indication of the proportional increase in value of dollars invested that results from bringing assets into the computer services industry. These ratios have a high rank correlation with one another and show no tendency to be best or worst for the largest or smallest firms. This is as expected if economies of scale are exhausted when firms are still small.

³³ See [64].

³⁴ Two firms were omitted because we had no market price data on them. One is privately held; the other is Canadian.

³⁵ The rank correlation with revenue and equity investment is high enough so that this size measure should not be misleading.

□ **Demand.** What is demanded of the computer services industry is information.³⁶ How much will be demanded depends on the amount and type of activity carried on in the nation, relative price, promotional and training effort, and the level of learning. There is some empirical evidence available on the effects activity, price, and learning have on industry demand. The impact of promotion and training and the distinction between industry demand and the demand for services from a particular firm are, however, matters illuminated only by discussion and unsystematic observation.

Evidence on the effects of activity comes from expenditure data for computer users. In using these data we assume that a change in the amount of activity will have the same effect on demand for information provided by computer services firms as it does on demand for all computer-based information. The parameters of the relationship between activity and usage of buyers can be estimated from a Diebold Research Program study of automatic data processing (computer) expenditure patterns for 245 large firms in 1971.³⁷ The Diebold conclusion is that “large companies do not systematically allocate a larger percentage of sales to ADP, nor do they tend to spend their ADP dollars significantly differently than smaller companies.”³⁸ The cross-sectional regressions shown in Table 9 use

TABLE 9
REGRESSIONS BASED ON INDUSTRY EXPENDITURE PATTERNS FOR MANAGEMENT INFORMATION SYSTEMS, 1971¹

VARIABLES IN LOGS	DEPENDENT VARIABLE	CONSTANT	SALES OF AVERAGE FIRM ²	AUTOMATIC DATA PROCESSING EXPENDITURES OF AVERAGE FIRM	R ²
YES	ADP EXPENDITURES	-6.293	+1.221 (5.781)		0.720
NO	<u>ADP EXPENDITURES</u> SALES	0.765	+0.000 (0.779)		0.045
NO	<u>EXTERNAL SERVICES</u> ADP EXPENDITURES	12.393	+0.002 (0.811)		0.048

¹ THE INDUSTRIES AND NUMBER OF FIRMS IN EACH ARE AEROSPACE (10), APPAREL (7), AUTOMOTIVE PRODUCTS (14), BUILDING MATERIALS (6), CHEMICALS (13), ELECTRONICS AND ELECTRICAL PRODUCTS (17), ENERGY (15), FOOD AND BEVERAGE (19), FOREST PRODUCTS AND PACKAGING (11), HOUSEHOLD AND PERSONAL PRODUCTS (14), INDUSTRIAL EQUIPMENT (11), INFORMATION PROCESSING (8), LEISURE AND EDUCATION (9), METALS (10) AND MULTICOMPANIES (16).

² NUMBERS BELOW THE REGRESSION COEFFICIENTS ARE t RATIOS.

SOURCE: [17].

average firm data for Diebold’s fifteen industrial and manufacturing industries. The first regression, run in logs, suggests that expenditures on automatic data processing increase more than proportionally with sales. The difference between 1.221 and 1.0 is not significant, however, as the second regression makes clear. The third regression relates expenditures for external services to sales. If external services,

³⁶ The basis of information demand, as for other goods and services, is the profit or utility of the buyer. Marschak [40] discusses the choice among information channels, Stigler [59] considers the decision on optimum quantity of information, and Hirschleifer [26] overviews the theory of information. Bower and Downes [7] describe a computer services industry customer’s decision.

³⁷ See [17].
³⁸ See [17], Abstract.

which are not broken down in the study, can be taken to represent computer services, then this regression suggests that the ratio of computer services to in-house computing does not vary with sales. This is in keeping with our assumption that activity volume will have the same effect on computer services as it does on total computing. The Diebold data suggest that the activity elasticity of demand is close to one.³⁹

Selwyn's data, which we discussed earlier, provide added support for this conclusion.⁴⁰ His data for 119 four-digit industries in 1966 can be fit cross-sectionally in the equation:

$$R = b_0 S^{b_1} C^{b_2} P^{b_3},$$

where R is total rent for all computers used in an industry, S is total value of shipments, C is the four-firm concentration ratio, and P is the number of individual establishments operated by the four largest firms. The regression coefficients are $b_1 = +1.08$, $b_2 = +0.453$, and $b_3 = -0.392$, all significant at the one-percent level. Taking total rent to represent the quantity of computer information demanded and holding industry structure constant, these results indicate an activity elasticity just above but not significantly different from one.⁴¹

Results from the Selwyn data demonstrate that industry structure as well as volume of activity influences the demand for computer services. Even with structure included, however, the equation above explains only 52 percent of the variance in computer expenditures among industries. Other aspects of activity help to explain the rest. The Diebold study finds differences in profitability and methods of controlling data among these.⁴² And a Belgian study argues that the complexity of the tasks performed by the firm is an important determinant.⁴³ Industry structure, profitability, methods of control, and task complexity are a start toward explaining differences in the demand for computer services among firms, but clearly there are

³⁹ The data available in the Diebold Research Program Document do not permit us to do much more than we have reported. Individual firm data are not presented. There are figures on ADP expenditure-sales ratios for large and small firms in each industry. In eight of the 15 industries the ratio was higher for the larger firms. This is consistent with the activity elasticity of one, or perhaps a bit more, that comes from the cross-industry regressions.

⁴⁰ See [54] and our discussion of scale economies above. In the earlier discussion we dealt with average rent per computer and number of computers. Here we simply combine the two and deal with total rent.

⁴¹ Selwyn assumed that activity elasticity of demand was one and used these data to talk about scale economies. We assume rents represent information service for reasons briefly noted in the text above and take it that we have the right dependent variable for a demand equation. When we use Selwyn's measure of total power of all computers used in an industry [54], pp. 77-79, rather than rent as the dependent variable $b_1 = +1.391$, $b_2 = +0.869$, and $b_3 = -0.655$. In this case b_1 is significantly greater than one. We feel the power measure is artificial and misleading, but even if it is correct, our estimate on activity elasticity would be changed only from one to 1.4.

⁴² In the Abstract, [17], the following conclusions are offered: "There is a high-order correlation between ADP expenditures and corporate profitability. In general, the more profitable a company is, the more likely it is to spend more of its revenues on data processing and to allocate its ADP resources differently The location of control over the data processing function seems to be closely related to expenditure patterns. Generally, those companies with centralized control allocate substantially less money to data processing than those with decentralized control."

⁴³ See [15].

other activity characteristics that are also important. We will not search out these influences. But we do want to emphasize that their existence means that price elasticity as well as other demand parameters will differ significantly among organizations and applications.

Volume and type of activity affect overall and individual demand for computer services; so do price and learning. The most impressive work on price and learning effects has been done by Gregory Chow.⁴⁴ Chow pictures the equilibrium long-run stock of computers that users demand as a function of price and activity:

$$Q_t^* = b_0 P_t^{b_1} GNP_t^{b_2},$$

where Q_t^* = quantity of computing power that is a target or equilibrium level for use under year t conditions, P_t = monthly rental for a unit of computing power relative to other prices, and GNP_t = gross national product or activity of users in real terms in year t . Adjustment to this equilibrium takes place by a process in which the growth rate of the computing stock reaches its maximum early and is fairly constant thereafter. The diminishing pull of the target is at first more than offset and then just offset by the increasing impact of experience with the growing stock. The relationship is

$$\log(Q_t/Q_{t-1}) = A_0(\log Q_t^* - \log Q_{t-1}),$$

where Q_t = actual quantity of computing power in stock in year t . Substituting the first equation into the second provides an equation free of the target quantity:

$$\begin{aligned} \log(Q_t/Q_{t-1}) = & A_0 \log b_0 + A_0 b_1 \log P_t \\ & + A_0 b_2 \log GNP_t - A_0 \log Q_{t-1}. \end{aligned}$$

To estimate this equation Chow requires measures of quantity and price. He assumes, as we noted earlier, that the rental charge for a computer system reflects its capacity to produce information. Given this assumption, Chow estimates information value coefficients for multiplication time, memory size, and retrieval speed from a pooled regression explaining, in terms of these characteristics, the rental charges of 82 computer models introduced from 1960 to 1965. These coefficients, applied to the characteristics of any computer model, provide an index of that model's capacity to produce information. The index for a model multiplied by the number of units in place in a year measures the contribution to information capacity. The sum of contributions from all models available in a year is Chow's measure of information capacity or quantity of computer power.

The price index for a year follows from the quantity measure. The average rental charge for each model introduced in a year is divided by the model's information capacity index. This provides a price per unit of information capacity for each of the year's new models. The arithmetic average of these prices divided by the GNP deflator is Chow's measure of the monthly rental or price for a unit of computing power for the year.

With these quantity and price measures for 1955 to 1965 Chow estimates his demand equation. GNP is dropped from the equation because of the high time-series correlation with the price variable. The results of the fit are $A_0 \log b_0 = 2.950$, $A_0 b_1 = -0.3637$

⁴⁴ See [27].

($t = -2.107$), and $-A_0 = -0.2526$ ($t = -3.417$). The R^2 is 0.834. This indicates a long-run price elasticity, b_1 , of -1.44 , a short-run price elasticity, $A_0 b_1$, of -0.3637 , and a combined learning and long-run price adjustment factor, A_0 , of 0.2526. With a 0.2526 adjustment factor the approach to target is about half-completed in three years. If *GNP* is included, and its close relationship to price handled by assuming that b_2 is twice b_1 , the only change of importance is a drop in the long-run price elasticity to -1.06 . With b_2 set at 1, as our cross-sectional work on activity volume would suggest, long-run and short-run price elasticity are virtually the same as in estimates excluding *GNP*.⁴⁵

The evidence from Chow's work and the studies of activity effects leave unanswered questions about demand differences peculiar to computer services, promotion and education effects, and distinctions between firm and industry demand functions. Because computer service is a closer substitute for in-house computing than computerized information is for all other information sources, Chow's price elasticity estimates should understate price elasticity in computer services. The difference may be quite small, however, since there are, as yet, only a small number of overlapping applications. Chow's work also provides some clue to the importance of promotion and education. Although he was unable to deal with these influences explicitly, the importance of his adjustment factor suggests they have a strong effect. IBM's history also seems to argue that promotion and education are important. As for distinctions between firm and industry demand, they should be typical of an industry structure that is competitive.

For the industry, demand is elastic in the long-run and inelastic in the short-run. For the firm the situation may be just the reverse. In spite of differentiation, it is clearly easier for an organization to shift from one computer services firm to another than it is to shift between noncomputing, in-house computing, and computer services. Price will direct new users to a particular firm and convince old users to transfer programs, move files, and try another source of external data. All these are short-run reactions. In the long-run, if the concept applies where prices are necessarily responsive, the firm that uses price to win business will find it difficult to hold business attracted both from within and from outside the industry. The demand for the firm is price elastic in the short-run and much less price elastic in the long-run.⁴⁶

This demand situation combined with the cost characteristics we

⁴⁵ In [8], pp. 3-25 to 3-30, we go further with the work Chow initiates. The price index is modified to recognize that there are old as well as new machines available in each year. In recognizing old computers in the price index it becomes apparent that they enjoy a rental premium—a characteristic also noted in Sharpe [56], pp. 350-356. The premium we associated with accumulated software and operating knowledge—the explicit elements of a learning effect. This learning effect contributes to capacity, and so we also modify the quantity variable. Our final results for long- and short-run price elasticity are very close to Chow's. With activity elasticity set at one and with the software-operating knowledge effects set at low and high extremes long-run price elasticity was between -1.80 and -1.41 and short-run price elasticity was between -0.48 and -0.46 .

⁴⁶ The long-run is a conceptual, not a chronological, period in which all other prices remain the same while complete adjustment takes place. Here we are really dealing with a time period and a response in price.

discussed above suggests some problems for firms when industry demand drops or industry expansion is too fast for demand growth. The temptation will be to cut prices substantially in order to attract business that will use capacity and add little to cost. The result, if many firms do it, will be very little more business and lower revenues for each of them because industry demand is inelastic in the short-run. Cooperative discipline would be an advantage for firms in this situation. The question is: Can they achieve it?

□ **Control of markets.** Where there is a leader with some power of retribution, a small and stable population of firms with an obvious, well-defined common interest, a barrier to intrusion, and/or an outside entity able to dictate conduct, collusive control of markets can develop and endure. In computer services the conditions that exist are not of this sort. Instead, the conditions work against "discipline" (or collusive control) and make competition difficult to avoid.⁴⁷

The first obstacle to discipline is the wide variation among firms in objectives, attitudes, constraints, and backgrounds. There are 99 firms in the remote sectors for which we have reliable information.⁴⁸ Only 32 of these firms originated as remote computer services operations. The others were established by computer manufacturers, other firms involved in electronic data processing, communications common carriers, aero-space firms, financial institutions, universities, and nonfinancial businesses. Most of the firms that originated elsewhere maintain a direct association with the parent firm and consider their actions with the interest of the parent in mind. When the parent is a computer manufacturer and supplies inputs to the industry or an aero-space firm and the affiliate's main customer, the parent's interest constrains decisions and affects objectives. The result is that common interest is a passing thing, elusive, and difficult to act on.

The second obstacle to discipline is suggested by the first. If organizations with such a variety of activities find themselves with equipment, software, data, consulting expertise, or customer contacts that make entry possible and inviting, then there must be a very large number of potential entrants. Since firms coming into the industry generally take less than nine months to go from formation to operation⁴⁹ and the equity investment required of a new firm seems to run less than \$1 million,⁵⁰ it seems clear that even if firms currently in the industry could find a basis for cooperative discipline, that basis would quickly be upset by new entrants.

A third set of obstacles to discipline is reflected in the mix of activities computer services firms carry on. Table 10 suggests one aspect of the mix. Firms are active simultaneously in time-sharing, remote job entry, and conventional service bureau operations. Only

⁴⁷ This is also the conclusion of the Federal Communications Commission, [70], Tentative Decision, paragraphs 14-23.

⁴⁸ This is not a consciously structured sample of the remote sectors, but includes all the firms for which we could get reliable information. The group of 99 accounts for 75 to 90 percent of remote revenues. Data for these firms comes from [3], [62], International Data Corporation's Computer Installation File, our own mail surveys of firms in 1971 and 1972, interviews with some of the firms, and a variety of other public sources. For fuller development of the material in this section and for more detailed tables see [8].

⁴⁹ See footnote 86 below.

⁵⁰ See Table 14 below.

TABLE 10

ACTIVITY MIX OF 99 FIRMS PROVIDING REMOTE COMPUTER SERVICES, 1971

FIRMS IN	NUMBER OF FIRMS
TIME-SHARING	73
REMOTE JOB ENTRY	53
ON-LINE DEDICATED	25
CONVENTIONAL SERVICE BUREAU	36
FIRMS IN ONLY ONE SECTOR	
TIME-SHARING	16
REMOTE JOB ENTRY	4
ON-LINE DEDICATED	18
TOTAL	38
FIRMS IN MORE THAN ONE SECTOR	
TIME-SHARING AND REMOTE JOB ENTRY	23
TIME-SHARING AND REMOTE JOB ENTRY AND CONVENTIONAL SERVICE BUREAU	21
OTHER COMBINATIONS	17
TOTAL	61

SOURCE: [3], [62]

firms in on-line dedicated applications restrict themselves to a single sector. These firms enjoy economies of specialization but have no protection against entry because skills and equipment similar to their own are used throughout the industry. Another aspect of mix is activity closely related to computer services. Ten of 99 remote firms carried on only computer services activities, but 50 also sold software, 38 did consulting, 23 contracted as facilities managers, and 14 offered education programs.⁵¹ A final aspect of mix involves service offerings that differ among firms. The various aspects of mix reflect several obstacles to discipline. First, sellers can shift among industry sectors even when buyers cannot. This means that if a small group of co-operative firms within a sector establishes an effective disciplinary arrangement, that arrangement will be quickly upset. Second, consulting firms, software houses, and facilities managers must share knowledge that permits them to move easily into computer services. This indicates one more group of ready entrants. Third, a variety of firms in and out of computer services can be combined in the information processing channel of any customer. With more entities involved or able to be involved, there are many more possibilities for substitution and much more of a cooperation problem. Fourth, firms in computer services are differentiated from one another not only by mix of activities but also by service characteristics.⁵² Differentiation may confer some minor element of monopoly power on a firm, but it also increases the difficulty of reaching and reviewing any arrangement to limit rivalry.

⁵¹ These data are drawn from [3], [62], and our mail surveys. Another way of viewing the mix is in terms of the relative importance of computer services operations as opposed to consulting, facilities management, etc. Our estimates are that 19 firms have revenues of less than 10 percent from all their computer operations, eight firms have 10 to 90 percent, and 64 have 90 percent or more. We have no estimate for eight firms.

⁵² A review of the service descriptions for firms in [62] makes this clear. The positive aspects of service differentiation are discussed in considering the public's view of industry performance, Section 5, below.

A fourth obstacle to discipline is the limited concentration in the industry. Limited concentration is reflected in the computer equipment figures of Table 11 and the revenue figures of Table 12.⁵³ No

TABLE 11
CONCENTRATION IN REMOTE COMPUTER SERVICES MEASURED IN
COMPUTER RENTAL PAYMENTS, 1971

	MONTHLY RENTAL ¹ (\$ THOUSAND)	PERCENT OF RENTALS FOR 99 REMOTE FIRMS ² (%)	PERCENT OF RENTALS FOR ALL COMPUTER SERVICES FIRMS ³ (%)
1. CONTROL DATA	1356	11.7	3.8
2. GENERAL ELECTRIC	1206	10.5	3.4
3. COMPUTER SCIENCES	680	5.9	1.9
4. UNIVERSITY COMPUTING	560	4.9	1.6
5. GRUMMAN	559	4.9	1.6
6. NATIONAL CSS	552	4.8	1.5
7. TYMSHARE	536	4.7	1.5
8. WESTINGHOUSE	279	2.4	0.8
9. INTERACTIVE DATA	276	2.4	0.8
10. BOEING	276	2.4	0.8
11. COM-SHARE	240	2.1	0.7
12. SCI-TEK	204	1.8	0.6
13. ALLEN-BABCOCK	198	1.7	0.6
14. ITT DATA SERVICES	183	1.6	0.5
15. PHILCO-FORD	180	1.6	0.5
16. CALL-A-COMPUTER	169	1.5	0.5
17. SERVICE BUREAU CORP.	168	1.5	0.5
18. NATIONAL CASH REGISTER	162	1.4	0.5
19. SYSTEMS DIMENSIONS, LTD.	155	1.3	0.4
20. MULTIPLE ACCESS GENERAL	152	1.3	0.4
21. ALL OTHERS	3312	29.5	77.1
SOURCE: DIRECT MAIL SURVEY OF FIRMS, 1971 AND 1972, [62], [28].			

firm has as much as a 5-percent share of the industry as a whole, and, although it is not demonstrated by these tables, there is some indication that concentration has been declining.⁵⁴ Concentration exists within individual sectors, but we have already discussed a number of obstacles in the way of making inferences on market control based on sector concentration. Those obstacles suggest that the remote sectors, with the exception of data banks, be considered together. Taken together, using figures from Tables 11 and 12, these remote sectors show

⁵³ The tables are developed from different sources and report different things. Some of the variation in the standings of firms in the two tables occurs because the firms have different revenue to equipment rental ratios. The cost discussion above should explain why this would be expected. Still, much of the variation is simply due to errors in estimation. These errors make individual rankings suspect, but they should not obscure the fact that both tables tell the same believable story about concentration.

⁵⁴ Estimates of 1970 and 1971 revenues by firm for the remote sectors are offered in [51], p. 159. The estimates for 1971 are consistent with Table 12 but not identical to it. In 1970 the share of all remote revenues attributed to General Electric was 14.5 percent and to the top 20 was 65 percent. In 1971 those shares were 12.8 percent and 61 percent. The same firms were included in the top 20 in each year, probably biasing the later figure downward.

TABLE 12

CONCENTRATION IN THE COMPUTER SERVICES INDUSTRY MEASURED BY ESTIMATED REVENUES, 1971

	TOTAL, COMPUTER SERVICES INDUSTRY		CONVENTIONAL SERVICE BUREAU		DATABANK		REMOTE JOB ENTRY		TIME SHARING		ON LINE DEDICATED TRANSACTIONS		UNDIS- TRIBUTED
	(\$MIL)	(%)	(\$MIL)	(%)	(\$MIL)	(%)	(\$MIL)	(%)	(\$MIL)	(%)	(\$ MIL)	(%)	
1 SERVICE BUREAU	70	4.7	40	3.8					20	10.8			10
2. GENERAL ELECTRIC	60	4.0							60	32.4			
3 AUT. DATA PROC.	45	3.0	45	4.3									
4. MCDONNELL-DOUGLAS	35	2.3											35
5. UNIVERSITY COMPUTING	33	2.2					20	20.0					13
6 CONTROL DATA	32	2.1					30	30.0					2
7. STAT TAB	30	2.0	30	2.9									
8. NATIONAL CASH REGISTER	25	1.7	15	1.4							10	18.2	
9. BUNKER RAMO	24	1.6			24	20.9							
10. TYMSHARE	15	1.0							15	8.1			
11 NATIONAL CSS	8	0.5							8	4.4			
ALL FIRMS	1495	100	1040	100	115	100	100	100	185	100	55	100	

SOURCE: [27].

the largest firm with 12 percent of equipment and 18 percent of revenue and the top four firms with 33 percent and 38 percent. Ratios like this would cause these sectors to be labeled unconcentrated by Kaysen and Turner even among their more inclusive industry groups.⁵⁵

The data bank sector requires special comment in connection with concentration. Within this sector particular data are often available from only one computer services firm. This is because the markets for many data banks are so small that they fail to exhaust even the limited scale economies that exist. Over-the-counter stock price quotations, for example, would cost more if the National Association of Security Dealers had franchised more than one firm to provide them. This is a case, and there are others, where there is a natural monopoly. It is also a case where Demsetz' question, "Why regulate?," is worth asking⁵⁶ because, although there is only one seller, there is the opportunity for many firms to bid to be that seller. The group selling the franchise usually creates the information in the first place, is knowledgeable, and finds greatest benefit in having the information distributed as well as possible at the lowest possible cost. The franchise can be for a short period⁵⁷ and involves resources, very few of which are irreversible. The conditions are ideal for franchise bidding to provide a competitive solution for a naturally monopolistic operation.

⁵⁵ See [35], pp. 26-37.

⁵⁶ See [16].

⁵⁷ Bunker Ramo's contract with the National Association of Security Dealers is for five years. The period seems about right. Software knowledge and machine capability progress enough in 60 months to make system redesign a cost-minimizing alternative. This short span for cost-minimizing contracts reduces the problems of contract administration and adjustment that make competitive bidding for utility franchises unworkable. It also means that errors in selecting a bidder can be corrected without suffering extensive damage and that a winning bidder's performance is being monitored constantly by those he serves and those he might serve on other contracts.

The four obstacles to discipline we have listed might be overcome if markets were limited geographically. In this industry, however, geographic limitations work against discipline. The costs associated with data transfer at one time were expected to cut the nation into a large number of small, isolated markets.⁵⁸ They have not. Of the 99 remote computer services firms eleven serve 26 or more states with IN-WATS toll-free service, and another six serve 26 or more cities toll free on direct private lines. These 17 firms do a national business. Thirty-eight firms have five or fewer states IN-WATS and five or fewer direct private lines. They do a local business. The remaining 44 firms, except for four on which we have inadequate information, do a regional business. The intersection of these firms and the resellers for them gives New York City 53 sellers, Chicago 46, Washington 33, and 26 other cities 10 or more. It provides every state but Alaska and Hawaii with five or more toll-free sellers of remote services.⁵⁹ No group of sellers in the industry is protected by geographic location.

The geographic distribution has an unexpected effect on industry discipline. The largest firms sell in many markets. Most of them sell from a single computer installation or a linked network of computers that offer the same service alternatives and the same price structure to customers regardless of the location at which they sign on the system. In each geographic market they deal with customers who think only of the alternatives available in that market, and they meet rivals who sell regionally or locally. The accommodations these national firms can make in any one market and the leadership they can provide are severely constrained by the conditions that exist in every other geographic market. This is not a situation that encourages cooperation or supports discipline, nor is it one in which cooperation and discipline would be easy to maintain.

A single supplier with control over key industry inputs might overcome the obstacles to collusion or, instead, take advantage of them to draw monopoly profits through the industry to the lower (input) level. There are two suppliers that deal with computer services that might do this. AT&T is one and IBM is the other. Later we discuss why and how each might move in this direction.⁶⁰ Here we simply want to note that although AT&T does provide virtually all communication lines used by the remote sectors of computer services, IBM does not provide all the computers. What they do provide is indicated in Table 13. IBM's share is 23 percent by number and 30 percent by value, well below its 60- to 70-percent share of the nation's total computer stock. The share figures in the table are for remote firms only. A 1970 survey by the ADAPSO estimated 39 percent by number for the computer services industry, software houses, and facilities managers.⁶¹ Adjusting this to a value share considering IBM's larger, more costly computers results in a 55-percent figure.

⁵⁸ Baran [4], p. 79, commented "These early general-purpose systems can, for the most part, only serve customers economically within 50 to 100 miles from the central computer. The economic limitation on market size is the high cost of long-distance telephone lines. We may, therefore, expect to see many small, independent 'computer utilities'—at first."

⁵⁹ All communications information is from [62] or its earlier editions.

⁶⁰ See the discussion of the supplier's view of performance in Section 5 below.

⁶¹ See [2], 1970 edition.

TABLE 13

COMPUTERS USED BY FIRMS PROVIDING REMOTE
COMPUTER SERVICES, 1971

COMPUTER MANUFACTURER	FIRMS USING ¹		TOTAL MONTHLY RENT ²	
	(NUMBER)	(%)	(\$000)	(%)
BURROUGHS	12	10	428	4
CONTROL DATA	10	8	2046	18
DIGITAL EQUIPMENT	16	13	430	4
HEWLETT-PACKARD	8	6	89	1
HONEYWELL (GE)	22	18	1907	17
IBM	30	24	3477	30
RCA	2	2	80	1
UNIVAC	8	6	1783	15
XEROX	11	9	1044	9
OTHER	5	4	227	2
TOTAL	124	100	11511	101

¹ THE NUMBER OF FIRMS SUMS TO 124 RATHER THAN 99, BECAUSE 10 FIRMS USED COMPUTERS FROM 2 MANUFACTURERS, 6 FROM 3 MANUFACTURERS, AND 1 FROM 4 MANUFACTURERS.

² THE PERCENT FIGURES SUM TO 101 BECAUSE OF ROUNDING. ESTIMATES OF AVERAGE MONTHLY RENTAL FOR COMPUTERS ARE IN [28], VOL. 6, NO. 9, MARCH 12, 1971.

SOURCE. [62], [3], DIRECT MAIL SURVEY OF FIRMS 1971 AND 1972.

The share is impressive, but, certainly in the remote sectors, it does not suggest that IBM now has control of the industry.

5. Four views of performance : public, seller, buyer, and supplier

■ The current structure of the computer services industry seems competitive. There is nothing in the cost or demand relationships and nothing in the disciplinary setting that should cause that to change. If there is going to be a change away from competition, it will have to be the result of action affecting structure initiated by some group dissatisfied with performance. Here we indicate how four interested groups see performance, why some are dissatisfied, and what the dissatisfied groups are doing to work for change.

□ **The public view.** In the public view, at least as economists conceive it, performance is not only satisfactory but also likely to be better than it would be with a different structure. Because the current structure is competitive, perhaps that can be taken for granted. Still, it seems worthwhile to consider service, price, return, innovation, resource movement, and side effects to see if there are any obvious shortcomings that might be correctable.

The "independent striving for patronage by sellers in a market," which is competition, results in a very large set of choices for the users of computer services. Sellers in the market differ in computers, delivery techniques, software characteristics, and pricing methods. The differences are clear in the group of 99 firms in on-line dedicated applications, remote job entry, and time-sharing.⁶²

⁶² This is the same group of 99 firms referred to in the preceding section. Ninety-five of these firms are listed in the September 1972 update of [62], and the information from that directory is the principal source for the material presented here.

Computers. The firms have in use 60 models of computers from ten manufacturers representing virtually the whole range of hardware alternatives.

Delivery. Although all of these firms are on-line sellers, they offer a wide range of delivery alternatives. The free delivery area ranges from local to national, with 16 firms offering only local delivery and 22 firms covering 16 or more areas with toll-free service. The method of providing free delivery differs. Six firms use only IN-WATS as a method of delivery beyond their local area and 43 use direct lines and no IN-WATS. The speed of delivery varies. Nineteen firms receive and deliver information only at conventional teletype speeds of 10 or 15 characters per second, while 27 offer up to 240 characters per second, nine offer up to 480, and three offer up to 960. The terminals from which users can send or receive information are of many types and include not only all variety of teletypes and typewriter devices but also high speed printers, plotters, video displays, card readers, touch-tone phones, and analog monitoring devices. Dependability of delivery is also a matter of user choice. Some sellers, by the amount and sophistication of their equipment and the redundancy of their routes, offer communication almost free of error or interruption, while others add no protection to what the common carriers provide. In terms of area, method, speed, terminal, and dependability, a user can choose the combination that fits his information needs.

Software. The software alternatives open to users are far greater than either the hardware or delivery alternatives. They involve languages, programs, data, and system characteristics. The *Time-Sharing Directory* lists 34 available languages.⁶³ One firm offers its customers 16 of them, and 19 firms offer 10 or more languages. Sixteen firms offer no languages and provide no programming facilities for the user. Instead they provide a finished, usually specialized, package of programs and data files. These programs and data files and those in the libraries of other firms range widely over industries and activities. They apply to egg industry optimization, information retrieval, portfolio pricing, text editing, daily record keeping for auto dealers, land surveying, savings and mortgage accounting, electronic circuit analysis, secondary school administration, photocomposition, electrocardiogram processing, general statistics, numeric control of machines, chemical process simulation, water utility billing, and many other things. The systems also offer variety in technical and administrative characteristics. This includes choices on file handling, string manipulation, significant digits, language translators, flowcharting, and debugging aids. It also means the user may choose a system that offers more security (passwords, scrambled files, specification of valid users, automatic recording of those who call particular files), more dependability (files put on tape within 30 minutes of change, independent hardware facilities with common access, return to last instruction after disconnect) or better administrative control (customer control of his organization's HELLO sequence, user

⁶³ See [62], Section VII.

numbers, use limits, invoices). Again, the opportunity exists for organizations to choose computer services offerings that closely match their information requirements.

Pricing. Different pricing systems as well as different prices for different services exist. There are firms that sell their services on the basis of transactions or programs run, others that charge a flat monthly fee for an open delivery line, and still others which charge for individual inputs such as central processing, memory, and printing. Among the firms that charge for individual inputs there are substantial variations in pricing system particulars, but initiation or minimum monthly fees and price premiums for prime time, rapid response, and higher speed lines are common.

This listing of more obvious differences among the offerings of computer services firms understates the variety of alternatives available from the industry by not including subtle differences in education efforts, consulting, tailoring library programs, and other labor services that also permit customers to distinguish among firms. The presence of extensive service differentiation is a proper response to the various information needs within and among organizations. The differences are quite real, and the buyers are knowledgeable. Differences are not in style and promotion, the characteristics that seem to dominate the discussion of product differentiation and give differentiation a bad name.⁶⁴ There is no doubt that an omniscient regulator could produce a net resource saving by reducing somewhat the differences that now exist. But, with omniscience unavailable and much to be learned about effectively meeting information needs, the degree of differentiation here seems more satisfactory than that which another, more disciplined structure would produce.⁶⁵

Let us turn now from services to prices. Prices have an optimal structure when they are equal to marginal cost or when their deviations from marginal cost are inversely proportional to demand elasticities.⁶⁶ Discovering prices that suit these requirements and demonstrating that they have been found are difficult. The variation and change of price structures found in computer services is part of the process of discovery, a natural characteristic of competitive

⁶⁴ Scherer [53], pp. 324-345, begins his chapter on differentiation by noting that diversity is beneficial and making clear that "economic theory has provided no operational calculus for comparing the benefits of diversity with its social costs." His chapter, however, deals primarily with image differentiation, adverse entry effects, and other problems. The chapters on the subject in recent books by Vernon [72], pp. 67-78, and Boyle [9], pp. 52-73, also leave the reader with the feeling that product differentiation is not really a good thing. Much of the feeling is due to the adverse effect on entry. It is appropriately paradoxical that a firm's ability to differentiate itself is one of the factors that makes entry so easy to achieve in computer services.

⁶⁵ Perhaps the industry approaches the ideal structure Chamberlain sought in [12], p. 215: "Differences in tastes, desires, incomes, and locations of buyers, and differences in the uses which they wish to make of commodities all indicate the need for variety and the necessity of substituting for the concept of a 'competitive ideal' an ideal involving both monopoly and competition."

⁶⁶ In [5] Baumol and Bradford discuss the question of optimal departures from marginal cost pricing and discuss as well three other variants of an optimal departure theorem. Although the variant we sight is less general than two of the others, requiring zero demand cross elasticities, it gets the basic idea across much more easily than the other alternatives.

search for something that is essentially unknown to any single firm in the system. To demonstrate whether the process is working, we refer to the discussions of cost and demand and note some specifics in the pricing of the 13 firms providing time-sharing from Honeywell G200 or G400 series computers.⁶⁷

Both hourly charges for connection of a customer with the system and central processing unit charges are greater on the higher rent, more powerful G400 systems. Users that transmit and receive at higher speeds pay more for each hour of connection because fewer can be serviced simultaneously by the communication interface. They pay no more for central processing unit time, which is unrelated to transmission speed. Other examples of prices that relate to long-run marginal cost and, in some cases, to demand elasticity are additional charges for specific programs or for software systems of greater capability, lower charges for off-line storage combined with fees for transferring data among storage devices, and fixed monthly fees for an open line. The lower charges for off-peak business, which one firm even brings to zero for processing, reflect the fact that cost varies little with off-peak use. Discounts to educational institutions are consistent with the smaller amount of programming and consulting they require and with their greater demand elasticity. Volume discounts and monthly minimums are appropriate because some costs of serving a customer are not associated with volume of computer use. Finally there is the processing charge, which for most firms now includes file access time, memory use, and the time involved in swapping in and out of the central processor.

We argue that prices provide users information on the resource requirements that go with their choices. This may not be a convincing demonstration that price structure approaches optimality. Similar pricing practices might be found with monopoly, although perhaps not under regulation.⁶⁸ The monopolist, however, need not be motivated to search out alternatives. There is evidence that these selling firms are searching for alternative buyers. Changes in price structure were noted for seven of the thirteen Honeywell G200 or G400 firms in comparing the *Time-Sharing Directory* of September 1972 with that of July 1970.⁶⁹ All of the changes moved in what seems to be the right direction, that is off-peak rates, volume discounts, transmission speed changes, and the like. The monopolist, if he did search, would turn his discoveries to excess profit. Firms in computer services have not. Their returns are in the range we might expect from competition.

Returns in this industry cannot really be found from accounting statements. The problems of distributing cost through time are just too great. An alternative is to calculate return using the amounts equity investors have put into a firm and the value the market places on the firm considering its practices and prospects. Table 14 shows

⁶⁷ The firms are all from the set of 99 described earlier and the data are from various editions of [62].

⁶⁸ See [50] for a view on this.

⁶⁹ The percentage was about 65 percent for all of our 99 firms appearing in the two directories and having the same equipment reported in both. The directories [62] come out at changing intervals, and two others appeared between these dates. Because firms are neither aggressive in updating the information they provide nor pushed to provide all of their pricing practices, the changes observed probably understate those that occurred.

TABLE 14

RETURN ON INVESTMENT (THOUSANDS OF DOLLARS)

	EQUITY INVESTMENT ¹					EQUITY INVESTED BY THE END OF FISCAL YEAR 1971	MARKET ² VALUE OF COMMON STOCK DECEMBER 1971
	1967	1968	1969	1970	1971		
ACADEMY (DEC.) ³	16	344	1907	430	0	2697	1553
APPLIED LOGIC (SEPT.)	171	1754	4975	16	65	6881	544
COMPUTER COMPLEX (JUNE)		3572	3377	0	52	7001	355
COM-SHARE (JUNE)		427	6774	690	885	8776	4323
INTERNATIONAL TS (MAY, t + 1)		2351	2730	15	162	5258	8973
KEYDATA (JULY)		2378	1016	4436	116	7946	15579
METRADATA (DEC.)			1826	0	47	1873	3696
NATIONAL CSS (FEB., t + 1)		699	4311	0	16	5026	8613
ON-LINE (APRIL, t + 1)	326	1466	(17)	15	280	2070	4924
PROGRAMS AND ANALYSIS (NOV.)			153	0	0	153	2760
RAPIDATA (DEC.)	204	286	16	313	400	1219	8843
TIME SHARE (OCT.)	259	75	1729	0	111	2174	821
TIME SHARING RES (MAY, t + 1)			748	0	0	748	1550
TYMSHARE (DEC.)	854	899	640	3606	1009	7008	24531
TOTAL	1830	14251	30085	9521	3443	58830	88055
RATE OF RETURN ON INVESTMENT 14 PERCENT ASSUMING INVESTMENT AT START OF YEAR							
RATE OF RETURN ON INVESTMENT 21 PERCENT ASSUMING INVESTMENT AT END OF YEAR.							
¹ DRAWN FROM FINANCIAL STATEMENTS, THIS IS OUR BEST ESTIMATE OF COMMON STOCKHOLDERS' CONTRIBUTIONS IN EACH YEAR. FOR SOME FIRMS, COMPUTER COMPLEX IS AN EXAMPLE, THE INVESTMENT SHOWN IN THE FIRST YEAR WE REPORT ACTUALLY WAS SPREAD OVER EARLIER YEARS. WE DO NOT HAVE THE DATA TO DISTRIBUTE IT PROPERLY.							
² THIS IS BASED ON A DECEMBER 1971 STOCK PRICE QUOTATION FROM THE <u>BANK AND QUOTATION RECORD</u> OR <u>NATIONAL MONTHLY STOCK SUMMARIES</u> . THE STOCK PRICE IS MULTIPLIED BY SHARES OUTSTANDING AT THE END OF THE 1971 FISCAL YEAR TO GET MARKET VALUE.							
³ MONTHS IN PARENTHESES ARE THE DATES ENDING THE FISCAL YEAR. A FIRM IS IN THE FISCAL YEAR 1971 IF ITS YEAR ENDS BETWEEN JUNE 1971 AND MAY 1972.							

investment and market value information for the fourteen publicly held U. S. firms for which we have data. Nine of these firms had market values for common stock that exceeded the equity investment. Five did not. These five had respectively lost 42, 92, 95, 51, and 38 percent of their investment. The fourteen as a group had a time-adjusted rate of return of 14 percent or 21 percent depending on the assumption of beginning or end of year investment. The true rate is, of course, somewhere between these two figures. And considering the distribution of fiscal years and the fact that investment shown in the first available year was actually spread over earlier years for some of these firms, it is probably not far from 14 percent. A rate in this neighborhood is certainly not much above the return that should be earned in an industry that is just emerging.⁷⁰

⁷⁰ Little support seems to be needed for a claim that a rate in the lower part of the 14- to 21-percent range is no more than reasonable for the 1967-1971 period. The Fisher-Lorie study [19] of long-run performance of New York Stock Exchange stocks indicated a return rate of about nine percent. Inflation expectations in the late 1960s and early 1970s should raise that to 13 or 14 percent. Even electric utilities usually have been granted 12 percent in recent years [61]. It might be argued that little of the return variation in an emerging industry is systematic (highly correlated with the market) and so can be diversified away in a portfolio of holdings. This, according to the capital asset pricing model [55], could make firms in computer services low risk. Nonsystematic risk does not seem to disappear entirely in tests of the model [32], however, and the systematic variation in computer services is probably quite large. Computer Sciences

The 14- to 21-percent return on all equity invested can be contrasted with the return received by buyers of the first public stock offerings of computer services firms. The experience of 23 firms in the industry, nine of them firms from the previous table, is shown in Table 15.⁷¹ In only one of the 23 cases was the stock price in Decem-

TABLE 15
EXPERIENCE WITH FIRST PUBLIC OFFERINGS

FIRM	TSIS# ¹	DESCRIPTION ²	DATE OF PUBLIC OFFERING ³	ISSUE PRICE ³	ISSUE AMOUNT (\$000) ³	PRICE DEC. 31, 1971 ⁴
ACADEMY	A1		2/6/69	5.00	1500	1 125
APPLIED LOGIC	A5		1/28/69	25.00	5600	0.250
COMMUNITY COMPUTER	C8		4/24/69	4.25	744	0 125
COMPUTER DYNAMICS	C11	RESELLER	3/20/69	7.50	1073	2.750
COMPUTER KNOWLEDGE	C22		1/13/70	6.00	1200	3 125
COMPUTER RESPONSE	C14	OUT OF REMOTE	9/18/69	6.00	1800	0.375
COMPUTER SHARING	C16	MERGED	4/23/69	10.00	4100	0 095
COMPUTER TASK GROUP	C36	RESELLER	8/7/69	3.00	390	1 250
COMSERV	C32	OUT OF REMOTE	6/17/71	3 00	600	2 000
DIALOG COMPUTING	D13	OUT OF REMOTE	1/22/70	15.00	5250	0.050
INFORMATION SCIENCE	I19		8/13/69	6.00	960	2.000
INFORMATION SYSTEMS	I5	RESELLER	5/14/69	6.00	1410	0.125
INTERMAC	I21	RESELLER	7/15/69	10.00	1000	0.000
INTERNATIONAL TELECOMPUTER	I12	OUT OF BUSINESS	6/3/69	12 00	2220	0 000
INTERNATIONAL TIMESHARING	I16		6/5/69	10.00	3000	4 500
INTERSTATE COMPUTER SERVICE	I22	RESELLER	11/10/69	4.00	440	1.750
KEYDATA	K2		4/7/70	11 00	2200	6.750
MANAGEMENT INFORMATION SYSTEM	M3	MERGED	2/26/69	6.25	875	5.750
METRIDATA	M7		4/29/69	12.50	1563	10.500
NATIONAL CSS	N5		1/8/60	14.00	3500	8.250
RAPIDATA	R1		9/1/71	10.00	1000	9.750
TIME SHARING RESOURCES	T14		1/16/70	7.50	750	2 000
TYMSHARE	T6		9/24/70	6.00	3600	8 500

¹ FROM [62] AND ITS EARLIER EDITIONS.

² WHERE NO DESCRIPTION APPEARS THE FIRM IS IN ON-LINE DEDICATED APPLICATIONS, REMOTE JOB ENTRY OR TIME-SHARING AND IS INCLUDED IN THE GROUP OF 99 FIRMS OPERATING IN 1971 THAT WE USE ELSEWHERE. RESELLERS ARE FIRMS THAT ACT AS EITHER MARKETING REPRESENTATIVES OR AS CONTRACTORS PLACING SOFTWARE ON A PRODUCERS COMPUTER. NO RESELLERS ARE INCLUDED IN OUR GROUP OF 99.

³ FROM [48].

⁴ FROM THE BANK AND QUOTATION RECORD OR NATIONAL MONTHLY STOCK SUMMARIES.

ber 1971 above the offering price. The reason Tables 14 and 15 are so different is that buyers of the public offerings were getting a very small percent of the shares while providing a very large percent of the equity investment. The original, prepublic investors in the firms were, in turn, reaping the rewards of their development efforts. A return to entrepreneurs in excess of the return on equity capital alone is appropriate and expected. A loss by those who provide only equity capital is neither appropriate nor expected. The explanation of the

has been in the related field of computer software and is now in computer services. Its stock price history is long enough to permit estimation, and our estimates are that its systematic risk or beta coefficient is about 2, or twice that of New York Stock Exchange listed firms as a group.

⁷¹ The group includes all firms listed in [48] and given an identifying number by Time-Sharing Information Services [62] with the exception of two firms that never actually entered computer services.

loss may be that buyers of the public offerings saw the industry's growth prospects and bright future but failed to recognize that, in a competitive industry where entry is easy, assets are unlikely to earn more than a normal return.

Normal returns are consistent with static efficiency but indicate nothing about innovation. To evaluate the pace of innovation requires an idea of the opportunities that exist, an indication of the proportion and type of opportunities being realized, and an estimate of the resources used in realizing them. None of these requirements can be satisfied directly, and our conclusions on innovative performance from the point of view of the public, consequently, are tentative.

Whether the current structure of the computer services industry should be expected to provide a more satisfactory pace of innovation than an alternative structure might be suggested by the general literature on innovation. But there is little help from this source. The proposition that an industry composed of fewer, larger firms has more incentive and advantage in innovation has not received empirical support.⁷² There is some indication of a threshold size for development efficiency,⁷³ some suggestion that entry barriers slow development and diffusion of innovations,⁷⁴ and a good deal of conflicting evidence on the effects of funds flows and product diversification.⁷⁵ About all that can be said is that the structure that is best for innovation will vary with the conditions that face an industry.

The relevant conditions for computer services can be summarized as opportunities, innovation characteristics, and appropriation possibilities. There is a rich lode of technological opportunities built up from invention and innovation in computer hardware, communications, and management science. The computer services' innovations to be mined from this lode are, for the most part, product rather than process innovations.⁷⁶ They involve very little hardware development, a great deal of ingenuity in combining hardware and discovering capabilities of the communications system, and a substantial effort to create software that makes sophisticated techniques helpful to unsophisticated users. These kinds of innovations have low cost and short life.⁷⁷ Developers appropriate advantage from them because

⁷² See [52] and [38], pp. 215-217.

⁷³ See [53], p. 353.

⁷⁴ See [37], p. 216.

⁷⁵ Contrast [52] and [20].

⁷⁶ Although combining information and information technology into services with decision relevance is a product innovation for computer services, it is a process innovation for customers. It is successful for computer services as a product only because it alters the choice/production process of customers.

⁷⁷ That relatively small outlays are required for innovation in this industry seems apparent. Software is a key element in innovation, and the small size of firms that specialize only in software suggests small outlay requirements. One estimate of the development cost of most commercial software is this: ". . . For packages in this price [\$2000 to \$20,000], the purchase price usually represents one-fifth to one-tenth of the total cost of developing an equivalent package," [25]. Another comes from our interview with the executives of Rapidata. They indicated that PLEA, their data retrieval language, went from an idea to an operating system in less than two months, on the efforts of not much more than one exceptional systems programmer. Even innovations that combine hardware, software, and communications do not demand outlays that require giant firms. Tymshare, for example, has been able to blend these elements in a sophisticated international data network, and its total equity investment in 1971 was only \$7 million.

imitation involves cost comparable to development and because an innovation is part of the pattern of improving service that holds customers. Even when they are not first to a particular innovation, developers do not entirely waste their effort; their costs of imitation and related innovation become less.⁷⁸ Investment in innovation, therefore, has the type of expected return distribution of more conventional investments even when a firm's portfolio of development projects is small.

With these technological opportunities, innovation characteristics, and appropriation possibilities, a computer services industry of many, small, differentiated rivals seems very likely to realize a larger proportion of all types of innovative opportunities than would a concentrated or regulated industry. There are no comparative observations to confirm that current structure does better in this respect, but there is some indication that it does well.⁷⁹ Expenditures on software development by computer services firms are substantial. In 1971 for the ten firms on which we report data in Table 5 the mean and median figure was about 9 percent of outlays or sales. This compares to less than 6 percent of sales spent by U. S. computer manufacturers to develop systems, utility programs, languages, and other standard software in recent years⁸⁰ and to 2 percent of sales spent by all industries on research and development in 1966.⁸¹ The expenditures lead to new products, imitation, and diffusion. The result is reflected in the service variety we referred to earlier and in continuous change. Between July 1970 and September 1971 new applications programs or systems changes were reported for 45 of 78 firms in the time-sharing, remote job entry, and dedicated applications sectors. During the same period, all but seven of these firms also made changes in the reception speed, coverage, or terminal com-

⁷⁸ Behind this set of appropriation conditions is the legal protection available on the types of innovation found in the industry. The Supreme Court's decision in *Benson and Tabbot* confirmed for all software what was suspected for most: it is unpatentable. See "High Court Denies Computer Patent for Programming," *New York Times*, November 21, 1972, pp. 1, 59. Copyright, although available for computer programs, serves no purpose because it makes the instructions in the program public and does not prevent others from using them as the basis for building similar programs. Only trade secrets, under the laws of the states covering torts and contracts, provide a measure of protection. The owner of a trade secret has no property right other than the right to prevent employees, competitors, or other parties from taking the secret from him in an unfair or improper way. For a discussion of trade secrets and software see [6].

⁷⁹ Although the comparisons are quite imperfect, it may be worth noting some judgments on innovation in communications and computers, two industries that provide key inputs to computer services and that combine to offer computer services technological opportunities. In telephone, Shepherd's verdict seems to be that regulation biased innovation toward maximizing "exclusivity of their system" and slowed the pace of innovation in some areas, particularly terminal equipment. See [57], pp. 95-96, 102-110. In computers we have discovered no judgment on the industry but IBM has been described as "not regarded professionally as a leading inventor or innovator," [58], p. 227, and has been estimated to get less innovation from its development effort than most of its competitors. See [24], pp. 69-93.

⁸⁰ The software expenditure figures for U.S. computer manufacturers are estimates provided to us by the International Data Corporation. For sales we used revenues from general purpose computers as reported in [28]. The figures in millions with software expenditure first are: 1967, 230, 4895; 1968, 260, 6910; 1969, 325, 7145; 1970, 390, 7645; 1971, 450, 8410.

⁸¹ See [53], p. 349.

patibility of their delivery systems.⁸² The changes are small, continuous, and what should be expected from the set of technological opportunities.

Not all of the industry innovations are small and without impact on innovation elsewhere. In creating systems software to suit their needs, firms like Applied Logic, Comshare, and Fulton National Bank have given computer systems manufactured by Digital Equipment Corporation, Xerox Data Systems, and General Electric added value and have led these corporations toward new software and hardware developments. They may, at the same time, have added to the pressure on IBM to create systems that lend themselves to interactive time-sharing. Other firms, Tymshare is an example, have discovered how to exploit technical and pricing characteristics of the communications system, and have provided the operational setting for realizing the foreign attachment opportunities opened up by the Carterfone decision.⁸³ The role computer services firms play in the developing data networks may also be important. The pace of innovation outside as well as within the industry may be affected beneficially by a computer services industry with a competitive structure.

If there is a question about the impact of current industry structure on innovative performance, it is not whether too few opportunities are being realized, but whether too many resources are being used to realize them. Certainly there is the possibility of duplication of effort and the threat that firms will achieve normal profits, with each investing more in development than is justified by the marginal social return on innovation. There is, however, a lively market in software, and buying from others is an alternative when the cost of innovation of imitation is too high for a firm. There is also a very large pool of opportunities and a rather unstructured, experimental problem of converting them into realized aids for information users. Waste may exist in innovation, but our tentative conclusion is that it would not be reduced, only shifted from spending too much to realizing too little, if structure were altered.

The question of waste also comes up in connection with resources entering and leaving the industry. There are no important barriers to entry. The same is true of exit. This condition, continuous innovation, the difficulty in anticipating what as well as how much will be demanded, and the essentially competitive failure of one firm to anticipate the entry or expansion of others, lead to a high turnover of firms in the industry. Listings of firms in the time-sharing, remote job entry, and dedicated applications sectors from successive editions of the *Time-Sharing Directory*⁸⁴ reveal the movement:⁸⁵

	1/69	1969	1/70	1970	1/71	1971	3/72	1972	9/72
Total Listed	131		157		135		127		128
Additions (+)		+30		+32		+27		+6	
Deletions (-)		- 7		-58		-32		-5	
Unaccounted for (±)		+ 3		+ 4		- 3		0.	

⁸² The firms are all those of our group of 99 that are reported on in the July 1970 and September 1972 volumes of [62].

⁸³ See [68].

⁸⁴ See [62] and earlier editions.

⁸⁵ The change in mood that is associated with the movement is documented in what people in and out of the industry say about it. *Business Week*, June 22,

Does this movement suggest destructive competition and waste? There is no reason to think that it does. Quite the contrary. The movement itself suggests that neither long entry lead time⁸⁶ nor irreversibility of capital or labor inputs, characteristics that make competition destructive, is present.⁸⁷ The rush of entry and expansion in 1968 and early 1969 did leave the industry with over capacity. Departures, return of computers, and growth in demand corrected this situation by the middle of 1971. Price cutting accompanied over-capacity.⁸⁸ With all hours off-peak hours, their cost functions and their individual demand curves caused firms to see an advantage in lower prices. The advantage eluded each firm because all of them sought it, but it did pass on to the public, which acted on prices that reflected only the resources required to provide additional service. The firms suffered financial losses, but real resources were not wasted.

To get some idea of the disposition of resources that went with movement in and out of the remote sectors of the industry we tried to trace the 240 firms that have been recognized with an identifying number by Time-Sharing Information Services, publishers of the *Time-Sharing Directory*. One hundred twenty-eight of these firms appear as active industry participants in the September 1972 edition of the *Directory*.⁸⁹ Of the 112 that have disappeared all but 47 were resellers, name changes, nonentrants, or untraceable. Twenty of the 47 were acquired, and the resources they brought to the industry are still there. Ten went out of business, and 17 continued in business but stopped producing remote computer services. These 27 sold their computers or returned them to manufacturers or leasing companies for use elsewhere. Labor moved within and outside the industry without losing the advantage of skills that had been learned. In some

1968, pp. 121-123, headlined its report: "Time sharing zooms through the ceiling. Staggering gains are coming to this sector of the computer industry after its early stumbles. But while growth is luck, competition is also fierce, and a sharp shake-out is in offing." In the report people in the industry were quoted: "'Our sales are doubling every month,' says the manager of one time-sharing service. 'The growth curve goes straight up,' says another. 'The industry is exploding,' exults a third." In *Computer World*, January 7, 1970, p. 1, the tone was different: "1970 is already shaping up as the shakeout year for the service bureaus. Industry experts have predicted it as being a fast and brutal process. It appears to be underway at this time."

⁸⁶ We have information on the date of formation and the date service was first offered commercially for 37 of our 99 remote sector firms. These all seem to be firms or divisions newly created to enter this industry. Twelve of the firms were in operation within six months of formation, 23 within nine months, 29 within 12 months, and 32 within 15 months.

⁸⁷ For an excellent discussion of destructive competition see [34], Volume 2, Chapter 5, particularly pp. 172-209.

⁸⁸ See "Price Cutting Was the Name of the Game in 1969." *Computer World*, January 28, 1970, p. 48.

⁸⁹ Of these 128 firms, 24 are resellers and 104 are producers operating their own computer systems. In the work on the remote sectors we report through almost all of this paper, we set the resellers aside in order to avoid double counting of capacity, services, or pricing policies. In our list of 99 producers are four that do not appear in the *Directory* [62] and 95 that do appear there. The nine firms listed in the *Directory* that we do not use include five firms that were new in the September 1972 *Directory* and came to our attention too late to follow up and four firms that we suspect may no longer offer remote services. Our follow ups, at least of current listings, indicate the *Directory* is quite accurate and quite inclusive.

cases the accumulated value in promotion and education was recognized by sale of the customer base; in others the software developed during the years of remote operation continued to be marketed. Still, many of these firms found that the labor and machine time that went to development were lost. This is neither waste nor a reflection of destructive competition. It is part of the investment required to find services of greatest value to users in an industry where that is a particular challenge.

The final performance consideration for the public is adequacy of industry protection of privacy.⁹⁰ When a data bank of personal information is collected, assembled, and used, the costs to the computer-catalogued individuals usually are not included in the prices paid by collectors, assemblers, or users.⁹¹ The loss of privacy is not reflected in the market, and, because it is not, current performance of the computer services industry is unsatisfactory.

Privacy has been described as “. . . the right to determine what information about oneself to share with others.”⁹² Its protection is not a new problem.⁹³ But the problem has been intensified by what the computer can do with data. The computer's ability to store a large number of observations, to search quickly, to screen effectively, and to monitor unobtrusively makes data that previously would have been protected by the technical inability to get it available on individuals. Technical inability no longer affords any protection for privacy.⁹⁴

Concern for this lost protection has concentrated quite properly on government data banks.⁹⁵ But the problem also extends to private data banks in the computer services industry. It has many aspects. One is error in collecting and recording data. A second is the location and presentation of incomplete data. A third is the use of data by an audience for which it was not intended. A fourth is the ac-

⁹⁰ For a brief, clear comment of this see [11], Volume 2, p. 25, and for something more see [10].

⁹¹ Arthur Miller [46], pp. 34, 41, suggests that the protection of privacy is inadequate: “. . . not enough is being done to insure that computerized data, either in their stored form or in transit, are any more immune from the intransitive activities of snoopers than private telephone conversations . . . Computer systems that handle personal information may inflict harm on a data subject . . . by introducing factual or contextual inaccuracies in the data that create an erroneous impression . . .”

⁹² Westin, “Civil Liberties and Computerized Data Systems,” p. 152 in [22].

⁹³ See [74].

⁹⁴ This is a theme of Miller's article [47] and is stressed by a Honeywell Vice President, Robert P. Henderson, in the Ervin Hearings [67], p. 470: “I feel that the phenomenal speed and efficiency of the computer have raised the process of data collection, storage, retrieval, and dissemination to the point where it will be easier to invade the privacy of our citizenry.”

⁹⁵ Senator Ervin's opening statement in the data bank hearings tells the story. He asserts: “These hearings were called because it is clear from the complaints being received by Congress that Americans in every walk of life are concerned about the growing collection of information about them which is none of the business of the collectors. They are concerned about the confidentiality and security of the information on them which is in the hands of those whose decisions can affect their lives for better or worse . . . They are concerned that this information is being automated or computerized without proper screening or controls. But above all, they are worried that the existing laws are no longer sufficient to protect the privacy of the individual against the ‘information power’ of government,” [67], p. 1.

cumulation of data, for example travel or accomodation activities in reservation systems or student activities with exercises or computer-assisted instruction in educational systems. A fifth, and one that may turn out to be particularly important, is conversion of innocent, available pieces of information through synthesis. This would involve taking bits of information that are available on an individual—club memberships, arrests, employers, residences, etc.—and combining the innocent parts into a damaging whole. Each of these aspects is reason for concern.

Because none of them is handled effectively by the industry in its current legal setting⁹⁶ and because all of them seem important, there is some impatience for action.⁹⁷ The suggestions for action include the creation of an independent regulatory commission, establishing standards of professional conduct supported by statutory requirements for data protection, industry associations or licensing, declaring a freeze on the collection of certain data, and providing a group of ombudsmen who have reference to a data bank on data banks.⁹⁸ There is no easy way to choose among these suggestions, but several of them have the disadvantage of offering a base from which limitations on entry and freedom to compete could develop. And none of them provides a sensitive instrument to discover the balance that “. . . must be struck between society’s need for information and the individual’s right to privacy.”⁹⁹ This balance probably can better be discovered by “pinpointed laws that will provide the tools for due process in each situation . . .” and that extend to individuals a right to review files that affect them.¹⁰⁰ These laws, although difficult to develop, could remedy the market failure apparent in the computer services industry’s side effect on privacy by forcing sellers and users to bear the cost when privacy is threatened.

After considering service, price, return, innovation, resource movements, and side effects from a public view, it is possible to note ways in which industry performance could be improved but difficult to argue that any substantial change in industry structure would be

⁹⁶ See [76], pp. 305–307, for the difficulties of treating the privacy problems associated with computers in American law. Westin prefaces his remarks in this section, p. 305, with the statement: “In many ways American law is in the worst possible shape to deal with information processing and privacy, much worse than the task of modernizing its concepts in the fields of physical and psychological surveillance.”

⁹⁷ As Miller sees it, “Time is a luxury personal privacy cannot afford and the glacial movement of legal doctrine is inappropriate for the problem at hand,” [46], p. 235. More recently, Westin and Baker [77] have surveyed the databank problem and found the threat slow to materialize.

⁹⁸ Many of these suggestions are specifically directed to computer information/privacy problems in government but none of them is restricted to government. Suggestions on a regulatory commission are included in testimony by Bigelow, Chairman of the Special Interest Group on Computers of the Society of the Association for Computing Machinery [67], p. 692. Ideas about establishing standards are suggested by Arthur R. Miller [46], pp. 238, 267–272. Miller also discusses freezing or controlling data collection [46], p. 236, [47], pp. 22–31, as does Neuborne on behalf of the American Civil Liberties Union [67], p. 49. The data bank on data banks is put forward by Westin [67], pp. 816–818.

⁹⁹ Statement of Elliot L. Richardson, while Secretary of Health, Education, and Welfare [67], p. 811.

¹⁰⁰ The quotation is from Westin’s testimony [67], p. 815. The right to review is emphasized by Neuborne [67], pp. 52–53, Westin [76], p. 168, and Miller [47], p. 31.

beneficial. If the structure is to change, the initiative and the advantage will be elsewhere.

□ **The seller view.** Initiative for changing industry structure can be expected from sellers of computer services. For them, most of the performance characteristics that satisfy the public represent outcomes that fall short of aspirations and reasons to work for change. The industry gives them less return than it should, asks more of them than is reasonable, and fails to provide a setting in which they can serve the public properly.

The return seems too low when compared to the most obvious standard, the computer hardware industry. In computer hardware the leading firms, firms which dominate an investment weighted industry average, find each equity dollar contributed worth six dollars in stock market value. This compares to a ratio of less than two to one for the computer services firms.¹⁰¹ The accounting income return on book equity provides an even more unpleasant comparison not only when the yardstick is computer hardware but also when it is some general average.¹⁰² Dissatisfaction with industry comparisons is increased by the fact that there are very great differences within the industry in return experience. This may be explained by the large amount of unsystematic or residual risk involved, risk that may not be relevant for portfolio holders but that seems significant and unfairly compensated by managers.¹⁰³ Finally, there is the experience of the stockholders who, although deluded when buying, find that they have generally had losses rather than gains on new issues of computer services stocks. Return, in the eyes of the sellers, is just not adequate. Not only is return inadequate but also life in the industry is too difficult. Managing a variety of services that never quite satisfy customers and being caught in the "perennial gale of creative destruction" that continually forces change in services and systems is not a relaxing activity. Building price schedules that recognize differences and bargaining in each special situation take thought and cause tension. Seeing prices cut, customers unmoved, and competitors cutting further is frustrating. Watching cash flows drop when the revenue effects of declining business and ineffective price cuts are unmatched by cost savings is frightening. Having firms enter when demand picks up after lasting out a period of pressure and failure is unfair. Life in the industry is simply more difficult than it should be.

The inadequate return, the difficult life, is the result of not having any discipline in the industry. Lack of discipline, in the seller's view, is also the problem in serving the public properly. The public should have computer services with maximum security, a minimum error, and complete protection of privacy. The inability to provide these things in the industry's current, undisciplined structure is a matter of dissatisfaction for the sellers.

¹⁰¹ Two larger firms listed on the New York Stock Exchange, Computer Sciences and University Computing, both of which have substantial activity in computer services along with other business, do not change this picture. Neither had a ratio above 2:1 at the end of 1971.

¹⁰² This is a function of the cost relationships discussed above and the stage of industry development. Explanation does not eliminate dissatisfaction, however.

¹⁰³ See footnote 70 above for comment and references on systematic and unsystematic risk.

Dissatisfaction, the result of a difference between sellers' aspirations and their realizations, provides pressure to change industry structure. The evidence of that pressure is just beginning to be visible. There is the start of a merger movement that could bring the number of firms down to the point where mutual interest would be obvious enough to encourage discipline.¹⁰⁴ ADAPSO, the industry trade association, has taken a position in opposition to industry entry by banks, universities, communications common carriers, and firms that do only "incremental marketing" of computer services.¹⁰⁵ It has also protested hardware discounts to educational and charitable institutions and government while suggesting them for computer services. These policies, whatever their motivation, all tend to limit entry, diversity, and outside alternatives. Efforts have been made to license or at least certify programmers.¹⁰⁶ And, with this, there has been a growing interest in establishing standards that would put a floor under the services offered by the industry.¹⁰⁷ Each of these actions would limit alternatives and impose a discipline that cannot be achieved without collective action and government support. Finally, there is a call for a regulatory agency "responsive to the computer industry."¹⁰⁸ The agency ultimately, if not initially, could help assure that the public was properly protected, that life in the industry was not more difficult than it should be, and that return was set at an appropriate level.

The sellers' efforts to change industry structure have an honest and innocent origin.¹⁰⁹ The aim is not to benefit those in the industry

¹⁰⁴ Tymshare may be the best illustration and most active leader in this movement to merge. Prior to 1972 it acquired Dial Data and Graphic Controls. In November 1972 it acquired the time-sharing operations of Computer Complex, in March 1973 it acquired Megsystems and agreed in principle to acquire Allen-Babcock Computing. Some of these acquisitions are of firms in financial difficulties but others, Graphic Controls and Allen-Babcock, are not. RAPIDATA has indicated a desire to grow through acquisitions and has sought unsuccessfully to acquire Metridata. Control Data, in its settlement with IBM, took over IBM's Service Bureau Corporation and brought together two of the ten largest operations in the computer services industry.

¹⁰⁵ See, "ADAPSO Expands Role in Fight Against 'Incremental Marketing,'" in *Computer World*, January 24, 1973, p. 23.

¹⁰⁶ See *Computer World* "ECPI Proposes Proficiency Exam," January 7, 1970, p. 25, "Privacy Commission Chairman Suggest Licensing Ban," November 11, 1970, p. 6, "Effort to License Programmers Fails," May 10, 1972, p. 1, and "1973 Objective: Let's Make DP a Recognized Profession," January 10, 1973, p. 7.

¹⁰⁷ See "Readers Applaud Call for Complete Set of Standards," *Computer World*, July 5, 1972, p. 6.

¹⁰⁸ "In the area of government regulation, Bernard Goldstein [ADAPSO President] said the issue was getting inequitable treatment from the various boards that regulate the industry . . . The Federal Reserve Board ruled for allowing banks to enter into the business. Its ruling, he said, was 'extremely permissive.' These were not isolated incidents, he said, and they showed the need to create a government agency responsive to the computer industry and the need for free enterprises in the services segment of the industry. The agency would not be regulatory in nature, he said, but would rather serve to set consistent government policy and to redress industry grievances," *Computer World*, January 31, 1973, p. 23.

¹⁰⁹ The picture of a computer services industry performing very well but threatened with structural change from within is also painted by Charles Warden, a Vice President of the time-sharing firm, Data Resources, Inc. He sees less honesty and innocence than we do: "Proposals to overcome these evils are usually

at the expense of those outside. It is to do better for everyone. We have already argued, however, that the public will not be served so well if structure is permitted to change.

□ **The buyer view.** Buyers of computer services are unlikely to protest changes of the type we have been discussing even though they and the public beyond them would be hurt by the changes. The limited interest buyers show in any input that is a relatively small part of the budget is part of the reason, but even more important is the tendency of buyers with minor involvement to ignore the system and to observe only the firms they face.

What the buyers see with this viewpoint is service that could be better. They see a level of security which may be as much as they are willing to pay for but which is not so much as they would like. They see programmers who respond imperfectly to their requests, software packages that have defects, sellers that win their business and then disappear from the industry, special purpose programs which require reworking if they are to be shifted from one system to another, prices that take on meaning only in terms of the cost and effectiveness of the information they make available, and standards of performance which are unspecified or inapplicable. Each of these imperfections appears to have a remedy in a disciplined industry that would assure high quality, uniformity, and simplicity.

Buyers mislead themselves in this way by failing to recognize the diversity of their requirements, the imperfections of alternative structures, and the sacrifices in service and price that a disciplined group of sellers will demand of them. But, because they are mislead, they can be expected to provide passive support rather than active opposition to structural changes initiated by sellers.

□ **The view of input suppliers.** Sellers, however, are not alone in initiating actions that will change industry structure. The computer services industry's main suppliers are two firms described somewhat imprecisely as "the largest industrial monopolist[s],"¹¹⁰ IBM and AT&T. Both of these firms may find the computer services industry, with its current structure, forcing a wedge between their aspirations for performance and realizations. The actions they take in response to this pressure will affect industry structure.

For IBM the computer services industry is a threat to effective price discrimination and, ultimately, a threat to any monopoly profits that can be drawn from a computer hardware industry disciplined by a solidly established leader. In the early years the threat was not apparent because the industry was extending computer services to uses and users that were difficult for a hardware manufacturer to reach directly. During these years the computer services industry represented for IBM a welcome force that worked to shift demand upward for computer hardware. More recently, remote computer services firms have turned to commercial as well as scien-

devices to establish barriers to entry to protect the current suppliers from the challenge of new suppliers with a better mousetrap," [73].

¹¹⁰ "As the largest industrial monopolist (second among American firms only to the AT&T monopoly of telephone service and production of telephone equipment), IBM inevitably faces difficulty in enforcing internal efficiency . . .," [58], p. 231.

tific uses. They have become resellers of computer power and other information inputs to the customers in IBM's basic markets. If this effort continues and expands, it will limit IBM's ability to discriminate among groups of customers. The computer services firms would, in this case, skim some of the cream.

Would this hurt IBM and widen the realization/aspiration gap enough to induce action? It seems likely that it would. The theoretical case demonstrating that a monopolist adds to profits by discriminating is well established. IBM's actions are consistent with the case.¹¹¹ The company has used conventional price discrimination with hours of use, model types, parts, peripherals, and customers. Much more important, both before and after unbundling, it has varied service among customers while charging the same prices.¹¹² How much this has added to profit we will not attempt to calculate. But because the discrimination is so finely tuned by service variation and because IBM seems to react strongly when competition beyond basic hardware promises to limit fine tuning, the effect on profit must be substantial. Certainly it must be substantial enough to deserve protective action.

IBM is capable of protective action. Its well conceived marketing and service policies helped to create the computer industry. But continuing domination of the industry is not simply the fallout from initial wisdom. The Department of Justice claims that the actions required to achieve this include predatory price discrimination, accumulating software to bar entry, introducing or announcing fighting computer models, and using discounts to dominate the educational market.¹¹³ These claims have not yet been proved; but consistent with them are the experience of RCA and General Electric when they challenged IBM.¹¹⁴ IBM's challenge to peripheral manufacturers,¹¹⁵ and the excess liquidity in the company's financial structure.¹¹⁶

¹¹¹ See [58] for a more complete discussion.

¹¹² Scherer [53], p. 253, defines price discrimination as "... the sale (or purchase) of different units of a good or service at price differentials not directly corresponding to differences in supply cost." This definition, he points out, includes "... the execution of transactions entailing different costs as identical prices"

¹¹³ *United States of America v. International Business Machine Corporation*, United States District Court for Southern District of New York, Filed January 17, 1969.

¹¹⁴ Shepherd [58], p. 224, remarks that "in the face of such extreme growth, IBM's retention of its high market share . . . is without parallel in modern industry." He also points out that "[a] series of major firms have had little success in challenging IBM's full-line position" Alan McAdams of Cornell's Graduate School of Business and Public Administration has analyzed IBM's response to RCA and General Electric in a paper not yet made generally available.

¹¹⁵ *Datamation*, February 1, 1971, p. 45, in a report headed "IBM's New Peripherals May Be Too Much Competition for Competitors" noted IBM actions from June through December 1970. They included introducing improved peripherals, lowering prices, adding to imitators' compatibility problems by placing an unintegrated control unit in peripherals, and suing competitors. *Datamation* quoted one observer: "IBM is trying to tell the compatible peripheral makers that if they play follow-the-leader too closely, it will undercut, outperform, and/or sue them."

¹¹⁶ The financial structure of IBM seems strangely liquid and free of debt. In 1972 the company had total assets of 10.8 billion dollars. Two-and-one-half billion was cash and marketable securities. Long-term debt was less than \$800 million, and cash, marketable securities, and receivables exceeded all debt and reserves by \$1 billion. The cash flow from operations in 1972 was 2.8 billion, less than 500 million of it committed to debt retirement and debt service. In 10 years

Action can be taken by the equipment suppliers against the computer services industry. Divestiture of the Service Bureau Corporation in the Control Data settlement¹¹⁷ is no signal that it will not be taken. Divestiture only means that IBM recognizes that a subsidiary, operating independently and exposed to separate scrutiny, could not acquire leadership or even hold its own in the computer services industry.¹¹⁸ The computer services industry requires a different approach. Price and service discrimination favoring final users over services firms may be part of it.¹¹⁹ A more important part may be the subtle linking of computers in the field to a larger computer at IBM. The central computer would help maintain and repair the satellites, program them, and even act as their data depository.¹²⁰ There could be other, more imaginative parts to the approach as well. Taken together these parts could slow the growth of the computer services industry, direct development away from uses in which IBM discriminates effectively, and intimidate any computer services firm with an eye for other markets.

It might be argued that IBM's action to preserve the ability to discriminate is in the public interest if the hardware monopoly would remain in any case. IBM is a discriminating monopolist, and it breaks its markets into many segments. Forced to maximize profits with one price for everyone, the company would have to choose a price that excluded some segments entirely and that therefore reduced output.¹²¹ This would represent a social loss. But if the computer

since 1963, there had not been one in which the cash flow from operations had fallen. The finance textbook [71], Chapter 7, suggests that the tax shelter of interest payments, if nothing else, makes this structure too safe to maximize shareholder wealth. Although the textbooks provide no justification, IBM management is probably neither irrational nor unconcerned with investors. The structure is probably just what it should be for wealth maximization when the asset to be protected is monopoly power and the ability to discriminate and when financial strength permits immediate, massive retaliation.

¹¹⁷ See "I.B.M. to Sell Unit to Control Data in Settling Suits," *New York Times*, January 16, 1973, p. 1.

¹¹⁸ Quantum Science Corporation figures [51], p. 159, give IBM's Service Bureau Corporation 5.5 percent of revenue from the industry's remote sectors in 1970 and 4.6 percent in 1971.

¹¹⁹ There is some indication this may already be occurring. *Computer World*, January 28, 1970, p. 23, reported that IBM was pricing its time-sharing software so that smaller users could install it in-house. IBM's expectation was that time-sharing might cease to exist as an outside operation. On March 1, 1972, *Computer World*, p. 23, noted an ADAPSO Hearing on industry grievances: "Thrust of the early testimony was that IBM had unfairly singled out the DP service industry as a special target for poorer service than it provided other users."

¹²⁰ *Datamation*, February 1, 1971, pp. 47-48, reported IBM's remote maintenance scheme for the 370 system: "... IBM is implementing a remote maintenance support network . . . There is no extra charge to users except for phone company fees for the data access arrangement necessary to run remote diagnostics for system/370 mainframes." ADAPSO's concern with IBM action has frequently been expressed. Its comment on the Control Data settlement recognized that excluding IBM as a computer services seller might not be enough: "Such exclusion should also prohibit the supply of any support activities which would, in effect, represent a 'back door' entry into the data processing services industry," [1].

¹²¹ The three degrees of price discrimination are best viewed as reference points in a continuum in which the groups given separate treatment go from two to that very large number in which every use for every user forms a separate group. When the number of groups is small, as it is at the third-degree reference point, it is appropriate to take Scherer's position that "... it is impossible to determine whether on balance . . . discrimination increases output and improves the

services industry develops from its current structure in a way that helps erode IBM's monopoly position as well as its ability to discriminate, then the case for protecting the structure of computer services and the pressure on IBM to change the structure become much stronger. With or without this, IBM is a supplier that can be expected to take actions that will affect the structure of the computer services industry.

AT&T's situation is not very different from IBM's. The computer services industry is an important part of a changing communications environment that could alter AT&T's position and its ability to discriminate. For AT&T, however, actions affecting the computer services industry may be motivated also by the opportunity the industry presents for recovering what may be lost elsewhere.¹²²

A series of Federal Communications Commission decisions has opened the way for the remote sectors of the computer services industry. Carterfone¹²³ permits interconnection of other than telephone company equipment with the switched network, encouraging development of better devices for sending and receiving data.¹²⁴ The specialized common carrier decision¹²⁵ introduces competition in data communication and promises lower input cost and a wider range of communication choices.¹²⁶ The outcome of the computer inquiry¹²⁷

allocation of resources," [53], p. 259. When the number of groups is very large, as it is in IBM's case, then it is quite certain that there will be some groups for whom the highest price at which any output will be bought will be below the marginal revenue for all other groups considered collectively. Discrimination will then increase output.

¹²² The position and length of our discussion of the communications/computer services relationship properly reflects our conclusion that the crucial public policy decisions for computer services are not in this area. Perhaps this is because so much has happened recently to settle the communications policy questions that were most troublesome and because computer services firms adapt so well. For an earlier discussion of the issues that goes into much more detail and that reflects concern which we feel is no longer warranted, see [42].

¹²³ See [68].

¹²⁴ Carterfone did not close the foreign attachment interconnection issue. There is still the question of protecting against attachments that would injure other users of the system. For the moment, this question is resolved by requiring an interface device provided, at a charge, by the telephone company. The device has been no problem, but, as we note below, the charge could become a problem. A search for attachment specifications that would make an interface device unnecessary is now being conducted by the Federal Communications Commission. See the statement of Dean Burch, Chairman of the FCC in [66], p. 51.

¹²⁵ See [69].

¹²⁶ Kahn [34], vol. 2, pp. 132-136, emphasizes the low cost and new alternatives associated with the Microwave Communication, Inc., service. AT&T's response to this is not unexpected. It has filed a new long line rate schedule that departs from uniform nationwide pricing and charges users of high capacity private lines less than low capacity users. See "AT&T Long Lines Files New Rates," *New York Times*, February 27, 1973, pp. 47, 54, and *Computer World*, March 7, 1973, pp. 1-2. That competition has forced this reduction in the cost of communications for computer services firms and that the reduction moves communications prices toward marginal cost seem clear from the statement of Richard R. Hough, President of the Long Lines Department: "This move is required to meet changing conditions, notably emerging competition by specialized common carriers providing private line communications services on selected routes competitive with ours. By relating rates more directly to our relevant costs, the restructuring we propose will permit us to recognize competition on the high density, low-cost routes these carriers seek to serve."

¹²⁷ See [70].

holds off regulation of the computer services industry and protects industry structure against the actions of the data processing subsidiaries of communications common carriers.¹²⁸ These decisions helped to establish and worked to preserve the computer services industry structure we have described. But there are still decisions to be made on the relationships among carriers, industry boundaries, resale of communications, and data tariffs that could change industry structure. AT&T's interest opposes that of the computer services industry in each of these areas.

Decisions that remain on the relationship between AT&T and the specialized common carriers will give meaning to AT&T's right to compete, the terms on which carriers can interconnect, and the opportunity for communications users to weave the lowest cost combination of lines and services from the offerings of several carriers. The decisions are important to AT&T and sellers of computer services. Computer services firms and their customers are interested in digital transmission among a small number of locations. Their needs match the specialized common carriers' offerings. If the specialized common carriers are free to compete and to interconnect, then the price of data communication service will move toward marginal cost. In this case neither communication systems of their own nor common carrier prices that discriminate to hold their business would give larger computer services firms an advantage. Without such an artificial advantage concentration and discipline are less likely to develop in computer services. Competition by the specialized carriers, whatever benefit it may have for computer services structure, is not attractive to AT&T. Even in this early stage, competition has forced departure from the politically and economically attractive policy of uniform nationwide pricing.¹²⁹ The responses that could be sought include governmentally imposed barriers to interconnection, and prohibitions against customers' weaving services from several carriers together.¹³⁰ A successful effort by AT&T to get such FCC responses could change the structure of computer services industry.

Decisions on the boundary between communication and computer services also remain to be made.¹³¹ The Federal Communications Commission indicated in the computer inquiry: "It is our position that where message-switching is offered as an integral part

¹²⁸ A U. S. Court of Appeals has struck down the FCC rule that communications common carriers cannot buy services from their data processing subsidiaries. "FCC Can't Bar Telephone-Telegraph Firms from Computer-Unit Dealings, Court Rules," *Wall Street Journal*, February 16, 1973, p. 9. This takes away a small element of protection against communications common carrier subsidiaries growing to dominate the computer services industry, but the element is so small it can be ignored. If communications firms are going to gain domination by diverting revenues from regulated business to computer services subsidiaries, then they can do it in less obvious, more effective ways than direct payments for services. In any case, exclusion of AT&T from computer services on the basis of an earlier consent decree [65], rather than the computer inquiry decision [70], makes the communications common carrier subsidiary threat to computer services industry structure a minor one.

¹²⁹ See footnote 126 above.

¹³⁰ This type of response to the earlier threat of microwave following the FCC's Above-890 decision is described by Kahn [34], pp. 129-132.

¹³¹ For a full discussion of the boundary question both as a general problem and at the communications/computer frontier see McKie [44, 45].

of and as an incidental feature of a package offering that is primarily data processing, there will be total regulatory forbearance with respect to the entire service whether offered by a common carrier or noncommon carrier . . ."¹³² This states a position on hybrid offerings; but because "integral part," "incidental feature," and "primarily data processing" will be defined only by individual cases yet to come before the Commission, it does not clearly establish the boundary between computer services and communications.¹³³ Computer services firms could find opportunities that involve more storing, forwarding, and switching than they do data transformation, and AT&T could find message-switching services, such as rerouting doctors' calls, that involve storing, referencing, and transforming data.¹³⁴ AT&T benefits when more computer activity comes within the boundary of communication. The Commission also may benefit because it can ask more in services it deems appropriate from a monopolist if the monopolist has more markets to exploit. If AT&T persuades the Commission to move the boundary further into computer services, then the structure of the computer services industry will change. There will be fewer opportunities for competition and fewer competitors. Discipline, therefore, may develop not only in the areas pulled into regulation but in the rest of the industry as well.

Resale of communications or line-sharing by computer services firms involves another decision still in the future.¹³⁵ Computer services firms often find it worthwhile to build within the telephone system a variety of networks of their own which offer the redundancy, error checks, noise monitoring, and efficient routing appropriate to their data-processing purposes.¹³⁶ These networks within the telephone system are useful to firms other than the originator and could be built and marketed by firms that employ computers for this purpose only and not for data processing. Firms like Tymshare already have customers for their network rather than their processing. The networks may have some scale economies even if computer services do not. If they cannot be shared or resold, then they may threaten the survival of smaller computer services firms. A choice by the Federal Communications Commission to maintain prohibitions on resale when developing practice forces explicit consideration of the issue could create conditions that would significantly change the computer services industry. Since prohibitions on resale are essential

¹³² Tentative Decision of the Commission, paragraph 41, [70].

¹³³ The FCC's intent to define the boundary on a case by case basis is indicated in the Final Decision of the Commission, paragraph 33, [70].

¹³⁴ Typical of the continuing interpenetration are Western Union's complaint against Graphic Scanning, "Western Union Corporation Seeks a Court Order on Graphic Scanning: It Contends That Growing Rival Is Performing Communications Jobs Requiring Federal License," *Wall Street Journal*, September 19, 1972, p. 12, and Incoterm's complaint against AT&T, "FCC Told AT&T's CRT May Violate Antitrust Laws," *Computer World*, February 21, 1973, pp. 1-2. The article quotes Incoterm as saying, "Because of the sophisticated nature and usage of the [Model 40] terminal, the [AT&T] service would be a remote access data processing service or hybrid data processing service."

¹³⁵ The Department of Justice raised the resale issue in the computer inquiry and argues that there was economic justification and legal precedent for permitting it. See response of the United States Department of Justice [70], pp. 39-51. The Commission acknowledged the issue but did not discuss it or take action on it. Tentative Decision of the Commission [70], paragraph 9.

¹³⁶ See [63].

to effective discrimination, AT&T can be expected to work for their retention.

The final decision that has importance for computer services relates to special data or Information Systems Access Line (ISAL) tariffs.¹³⁷ Users of remote services require either a moment of connection for exchanging signals on information such as credit worthiness or a very long connection to write a program or transmit a substantial amount of data. These types of requirements differ from the pattern expected of conventional telephone system subscribers. Rate structures, which are based on expected patterns, just do not fit the computer services users. The result may be that, particularly in local loops, computer services firms use both more switching equipment and more lines than they pay for.¹³⁸ An appropriate remedy would be to change the rate structure to recognize the inputs actually called into use. The response of local AT&T affiliates has been a special data tariff. Although filed, no special data tariff has been approved.

The idea will not disappear however. A data tariff could be a simple way to approximate price to cost for computer communications services. Also, the filing of a special data tariff might allow opportunities for specifying a particular and unique price/cost margin for communications services for the computer users alone. To AT&T, under competitive pressure at so many points, the specific price/cost margin here might most likely be a lower price/cost margin than elsewhere (particularly in the light of the entry of specialized communications carriers working entirely in data transmission). But it is impossible to predict what the price/cost configuration is going to be at this time.

The issue is important, because a higher price/cost margin here and no lower ones elsewhere would mean that a monopoly profit was being taken on communications transmission. In that sense, competition in the direct computer services would not be able to deliver all of its benefits to the public, since the noncompetitive profit was being taken at another level of the industry. At this point, in the face of entry and higher costs for communications services to computer users, this seems very unlikely. But it should be a consideration in dealing with special tariffs before the state and national regulatory commissions.

6. Policy implications

■ Although this paper is intended as a policy guide, it would not be appropriate to end with an extended discussion of specific policy proposals. That would be inconsistent with our theme that efforts to change structure from within an industry are the natural response to good public performance. It would not be in keeping with our argument that the current structure of the computer services industry cannot be improved upon. And it would give an incorrect impression

¹³⁷ Tariffs of this type have been proposed in Ohio, Illinois, and California, but the proposals have been withdrawn or put aside for further study. In the Ohio proposal, which was instituted by the affiliate and encouraged by AT&T, rates for some data users would have tripled. The Ohio proposal was withdrawn because it was holding up action on a general rate increase. See *Computer World*, September 17, 1969, p. 9, and March 18, 1970, p. 1.

¹³⁸ See the article by William H. Melody, "ISAL Proposals Could Lead to Total Phone Repricing," *Computer World*, July 28, 1971, Supplement, p. 6.

that initiative rather than prompt and proper response is the principal concern in public policy.

Prompt and proper response can support the current structure of the computer services industry and permit the industry to develop in the most beneficial way. But what the response should be depends on just what forms are taken by the pressures for change. Whatever the forms, it seems unlikely that the response will be limited to a single policy agent. Instead, many may become involved.

Minor mergers within the industry that could ultimately turn competition into disciplined oligopoly may call for antitrust reaction; so may efforts by IBM or AT&T to bring some of the industry into their own sphere of operation and more obvious actions by IBM to intimidate or control the industry. Obstructions placed in the way of specialized common carriers, efforts to keep computer services within narrow boundaries, arguments against resale of communications, and attempts to achieve monopoly prices with data tariffs call for response from the Federal Communications Commission and/or the state regulatory bodies. Proposals on privacy and software protection that would restrict industry entry require defensive analysis by the Congress and state legislatures. Executive agencies as well as legislatures and regulatory bodies will have to respond to pleas for standards, certification, or licensing and to calls for special treatment in government procurement or exclusion of "unfair" competitors from the industry. And all policy agents should have some response to an effort intended to provide computer services with its own regulatory body.

The computer services industry is unique in only one respect. It achieved full-blown status as an industry at such an early age that it seemed to be necessary to advance into industry regulation very rapidly or "before it is too late." By responding to these pressures for change with policies and procedures for maintaining open markets, we can maintain a competitive industry structure. We should do so.

References

1. ASSOCIATION OF DATA PROCESSING SERVICE ORGANIZATIONS, INC. *Annual Industry Study*, 1969-1970.
2. ASSOCIATION OF DATA PROCESSING SERVICE ORGANIZATIONS, INC., DATA CENTER SECTION. *Statement of Position*. February 2, 1973.
3. AUERBACH INFO, INC. *Auerbach Time-Sharing Reports*. 1971.
4. BARAN, P. "The Future Computer Utility." *The Public Interest*, Vol. 8 (Summer 1967), pp. 75-87.
5. BAUMOL, W. J. AND BRADFORD, D. F. "Optimal Departures from Marginal Cost Pricing." *The American Economic Review*, Vol. 60, No. 3 (June 1970), pp. 265-283.
6. BENDER, D. "Trade Secret Protection of Software." *The George Washington Law Review*, Vol. 38 (July 1970), pp. 909-957.
7. BOWER, R. S. AND DOWNES, D. H. "The Time-Sharing Decision in Banking." *Journal of Bank Research*, Vol. 2 (Autumn 1971), pp. 9-21.
8. ——— AND MENGE, J. A. *Regulation, Competition and the Computer Services Industry*. Preliminary draft submitted to the Brookings Institution, October 2, 1972.
9. BOYLE, S. E. *Industrial Organization: An Empirical Approach*. New York: Holt, Rinehart and Winston, 1972.
10. BROWN, I. A., JR. "The Computer and the Law of Torts and Contracts," in Grace W. Holmes and Craig H. Norville, eds., *The Law of Computers*, Ann Arbor: Institute of Continuing Legal Education, 1971.

11. CANADIAN COMPUTER/COMMUNICATION TASK FORCE. *Branching Out*, Volumes 1 and 2. Ottawa: Department of Communications, May 1972.
12. CHAMBERLAIN, E. H. *The Theory of Monopolistic Competition* (6th ed.). Cambridge: Harvard Univ. Press, 1948.
13. CHOW, G. C. "Technological Change and the Demand for Computers." *The American Economic Review*, Vol. 57, No. 5 (December 1967), pp. 1117-1130.
14. CONFERENCE BOARD. *Information Technology: Some Critical Implications for Decision Makers*. New York: The Conference Board, Inc., 1972.
15. DEBRABANDER, B., DESCHOOLMEESTER, D., LEYDER, R., AND VANLOMMEL, E. "The Effect of Task Volume and Complexity upon Computer Use." *Journal of Business*, Vol. 45, No. 1 (January 1972), pp. 56-84.
16. DEMSETZ, H. "Why Regulate Utilities?" *Journal of Law and Economics*, Vol. 11 (April 1968), pp. 55-66.
17. DIEBOLD RESEARCH PROGRAM. *Expenditure Patterns for Management Information Systems—1971*. New York: The Diebold Group, Inc., 1971.
18. ———. *Management Opportunities and Objectives Studies: Data Communications in American Industry*. New York: The Diebold Group, Inc., 1972.
19. FISHER, L. AND LORIE, J. H. "Rates of Return on Investment in Common Stocks: The Year-by-Year Record, 1926-65." *Journal of Business*, Vol. 41 (July 1968), pp. 291-316.
20. GRABOWSKI, H. G. "The Determinants of Industrial Research and Development: A Study of the Chemical, Drug, and Petroleum Industries." *Journal of Political Economy*, Vol. 76 (March/April 1968), pp. 292-306.
21. GREENBERGER, M., ed. *Computers and the World of the Future*. Cambridge: M.I.T. Press, 1964.
22. ———. *Computers, Communications, and the Public Interest*. Baltimore: Johns Hopkins Press, 1971.
23. HALDI, J. AND WHITCOMB, D. "Economies of Scale in Industrial Plants." *Journal of Political Economy*, Vol. 75 (August 1967), pp. 373-385.
24. HARMON, A. J. *The International Computer Industry: Innovation and Comparative Advantage*. Cambridge: Harvard Univ. Press, 1971.
25. HEAD, R. V. AND LINICK, E. F. "Software Package Acquisition." *Datamation*, Vol. 14, (October 1968), pp. 22-27.
26. HIRSHLEIFER, J. "Where Are We in the Theory of Information?" *The American Economic Review*, Vol. 63, No. 2 (May 1973), pp. 31-39.
27. INTERNATIONAL DATA CORPORATION. *Computer Services*. Boston, 1971.
28. ———. *EDP Industry Report*, Vol. 7, No. 10 (March 30, 1972).
29. ——— AND ASSOCIATION OF DATA PROCESSING SERVICE ORGANIZATIONS. *Annual Industry Survey*. Newtonville, Mass.: 1971 and 1972.
30. IRWIN, M. R. "The Computer Utility: Competition or Regulation." *Yale Law Journal*, Vol. 76 (June 1967), pp. 1299-1320.
31. ———. "The Computer Utility: Market Entry in Search of Public Policy." *Journal of Industrial Economics*, Vol. 62 (July 1969), pp. 239-252.
32. JENSEN, M. C. *Studies in the Theory of Capital Markets*. New York: Praeger, 1972.
33. JOHNSTON, J. *Statistical Cost Analysis*. New York: McGraw-Hill, 1960.
34. KAHN, A. E. *The Economics of Regulation: Principles and Institutions*, Volumes 1 and 2. New York: Wiley, 1970, 1971.
35. KAYSER, C. AND TURNER, D. F. *Antitrust Policy*. Cambridge: Harvard Univ. Press, 1959.
36. KEYDATA CORPORATION. "Computer Characteristics Review." *Computers and Automation*, Vol. 20, No. 6B (June 30, 1971), pp. 24-54.
37. LITTLE, ARTHUR D., INC. *The Computer Industry 1970-1975*. Boston, 1971.
38. MANSFIELD, E. *The Economics of Technical Change*. New York: Norton, 1968.
39. MARKHAM, J. AND PAPANKE, G. F. *Industrial Organization and Economic Development: In Honor of E. S. Mason*. Boston: Houghton Mifflin, 1970.
40. MARSCHAK, J. "Economics of Information Systems." *Journal of the American Statistical Association*, Vol. 66 (March 1971), pp. 192-219.
41. MARTIN, J. *Telecommunications and the Computer*. Englewood Cliffs, N. J.: Prentice-Hall, 1969.
42. MATHISON, S. L. AND WALKER, P. M. *Computers and Telecommunications: Issues in Public Policy*. Englewood Cliffs, N. J.: Prentice-Hall, 1970.

43. MCGEE, V. E. AND CARLETON, W. T. "Piecewise Regression." *Journal of the American Statistical Association*, Vol. 65, No. 331 (September 1970), pp. 1109-1124.
44. MCKIE, J. W. "Natural Monopoly and Its Limits: Data Communications as an Illustrative Case." Paper presented at the Seminar on Problems of Regulation and Public Utilities, Dartmouth College, September 2, 1971.
45. ———. "Regulation and the Free Market: The Problem of Boundaries." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 1 (Spring 1970), pp. 6-27.
46. MILLER, A. R. *The Assault on Privacy*. New York: Signet, New American Library, 1972.
47. ———. "Computer Technology and Personal Privacy," in Grace W. Holmes and Craig H. Norville, eds. *The Law of Computers*, Ann Arbor: Institute of Continuing Legal Education, 1971.
48. *New Issue Performance Directory 1969 and 1970-71*. New York: New Issue Outlook, Inc., 1970, 1972.
49. PHILLIPS, A. *Technology and Market Structure*. Lexington: Heath-Lexington, 1971.
50. POSNER, R. A., "Taxation by Regulation." *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 22-50.
51. QUANTUM SCIENCE CORPORATION. *Computer Sciences Strategy Program: Final Report*. Palo Alto, 1972.
52. SCHERER, F. M. "Firm Size, Market Structure, Opportunity, and the Output of Patented Inventions." *The American Economic Review*, Vol. 55, No. 5 (December 1965), pp. 1097-1125.
53. ———. *Industrial Market Structure and Economic Performance*. Chicago: Rand McNally, 1970.
54. SELWYN, L. L. "Economies of Scale in Computer Use: Initial Tests and Implications for the Computer Utility." Unpublished Ph.D. dissertation, M.I.T., June 1969.
55. SHARPE, W. F. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance*, Vol. 19, No. 4 (September 1964), pp. 425-442.
56. ———. *The Economics of Computers*. New York: Columbia Univ. Press, 1969.
57. SHEPHERD, W. G. "The Competitive Margin in Communications," pp. 86-122, in William M. Capron, ed., *Technological Change in Regulated Industries*, Washington, D.C.: The Brookings Institution, 1971.
58. ———. *Market Power and Economic Welfare. An Introduction*. New York: Random House, 1970.
59. STIGLER, G. J. "The Economics of Information." *Journal of Political Economy*, Vol. 69 (June 1961), pp. 213-225.
60. ———. "The Economics of Scale." *The Journal of Law and Economics*, Vol. 1, No. 2 (October 1958), pp. 54-71.
61. SWIDLER, J. "Rate Regulation in New York State." *Public Utilities Fortnightly*, Vol. 91 (March 29, 1973), pp. 45-49.
62. TIME-SHARING INFORMATION SERVICES, INC. *Time-Sharing Directory: Volume 1, Industry*. Philadelphia, 1972.
63. ———. *Time-Sharing Today*. Vol. 3, Nos. 1 and 2 (May 1972).
64. ———. *Time-Sharing Today*. Vol. 2, Nos. 3 and 4 (July 1971).
65. *United States of America v. Western Electric Company, Inc., and American Telephone and Telegraph Co. Final Judgment*. Civil Action No. 17-49, U. S. District Court, District of New Jersey, January 24, 1956.
66. U. S. CONGRESS, SENATE, COMMITTEE ON COMMERCE. *Overview of the Federal Communications Commission*, 93rd Congress, 1st Session, 1973.
67. U. S. CONGRESS, SENATE, COMMITTEE ON THE JUDICIARY, SUB-COMMITTEE ON CONSTITUTIONAL RIGHTS. *Hearings on Federal Data Banks, Computers and the Bill of Rights*. 92nd Congress, 1st Session, 1971.
68. U. S. FEDERAL COMMUNICATIONS COMMISSION. *In the Matter of Use of the Carterfone Device in Message Toll Telephone Service*, Docket No. 16942, and *In the Matter of Thomas F. Carter and Carter Electronics Corporation, Dallas, Texas, Complainants vs. American Telephone and Telegraph Company, Associated Bell System Companies, Southwestern Bell Telephone Company and*

- General Telephone Company of the Southwest, Defendants. Decision.* June 26, 1968. Docket No. 17073.
69. ———. *In the Matter of Establishment of Policies and Procedures for Consideration of Applications to Provide Specialized Common Carrier Services in the Domestic Public Point-to-Point Microwave Radio Service and Proposed Amendments to Parts 21, 43 and 61 of the Commission's Rules.* Docket No. 18920.
70. ———. *Regulatory and Policy Problems Presented by the Interdependence of Computer and Communication Services and Facilities.* Docket No. 16979.
71. VAN HORNE, J. C. *Financial Management and Policy* (2nd ed.). Englewood Cliffs, N. J.: Prentice-Hall, 1971.
72. VERNON, J. M. *Market Structure and Industrial Performance: A Review of Statistical Findings.* Boston: Allyn and Bacon, 1972.
73. WARDEN, C. *Economic Policy of University Computer and Computer Network Services.* Mimeo. Remarks made before the NSF/EDUCOM Airlie House Conference on Computer Networks, December 10-12, 1972.
74. WARREN, S. D. AND BRANDEIS, L. D. "The Right to Privacy." *Harvard Law Review*, Vol. 4 (1890), pp. 193-220.
75. WEAVER, W. "High Court Denies Computer Patent for Programming." *New York Times*, Vol. 122 (November 21, 1972), pp. 1, 59.
76. WESTIN, A. F. "Civil Liberties Issues in Public Data Banks," in A. F. Westin, ed., *Information Technology in a Democracy*, Cambridge: Harvard Univ. Press, 1971.
77. ——— AND BAKER, M. A. *Databanks in a Free Society.* Chicago: Quadrangle Books, 1972.

SYMPOSIUM ON THE NEW URBAN ECONOMICS

- 593 Notes on the new urban economics
E. S. MILLS AND J. MacKINNON
- 602 Congestion cost and the use of land for streets
R. M. SOLOW
- 619 Optimum vs. equilibrium land use pattern and congestion toll
Y. ORON, D. PINES, AND E. SHESHINSKI
- 637 The optimum factory town
A. DIXIT

Notes on the new urban economics

Edwin S. Mills

Professor of Economics
Princeton University

and

James MacKinnon

Assistant Professor of Economics
Princeton University

Since about 1970, ten or twenty papers have been written making use of control theory or programming to analyze optimum or market equilibrium urban land use. The purpose of these notes is to survey and evaluate this new approach to urban economics. Most contributions to the new urban economics assume that employment is concentrated at the urban center, but that housing production functions permit the amount of housing per unit of land to vary with economic conditions. An optimum or equilibrium pattern of housing density is deduced as a function of distance from the center. All contributions include assumptions about the urban transportation system, and some have congestion built into the model. The last part of the paper is a discussion of discrete and continuous representation of urban space. The two possibilities lead to different mathematical representations, and we suggest that discrete representations may be more useful for many purposes.

■ “All theory depends on assumptions which are not quite true. That is what makes it theory.”

Robert Solow

1. Introduction

Urban model building has a considerable history. Throughout the 1960s urban specialists built models of the spatial organization and growth of urban areas. A good survey is the work of Brown *et al.*¹ These models had various purposes, but mostly they were intended to assist governments in planning the provision of public services and in planning land use regulation. Most contained considerable detail and many nonlinear relationships, and were solved on computers by ad hoc methods. Most were the work of noneconomists and contained assumptions economists find implausible. However, the most recent and most sophisticated model in this tradition was built at the National Bureau of Economic Research and is the work of economists.²

Edwin S. Mills received the A.B. degree in economics from Brown University (1951) and the Ph.D. in economics from the University of Birmingham, England (1956). His current research in urban economics is directed toward urban land use.

James MacKinnon received the B.A. from York University in 1971. He is now a Ph.D. candidate in economics at Princeton University, where his research is in modeling general equilibrium systems.

The authors are indebted to James Ohls, Michael Rothschild, and Robert Solow for comments on an earlier draft.

¹ In [3].

² See [5].

In the 1970s, a new genre of mathematical urban models, almost unrelated to the earlier tradition, has appeared. The new models are the subject of these notes and constitute the new urban economics in our title. The hallmark of the new urban economics is the use of fairly sophisticated mathematics—calculus of variations, programming and control theory—to characterize some fundamental aspects of urban structure. Incidentally, the new game is played by almost entirely new players. They are mostly general economic theorists who have recently turned their attention to the urban economy. The old and still thriving game is played mostly by people with longstanding professional interest in urban issues.

The new urban economics is a small but flourishing business. There are perhaps a baker's dozen of published papers, most notably the three other papers in this symposium, four papers in the March 1972 issue of the *Swedish Journal of Economics*, and several papers in various recent issues of the *Journal of Economic Theory*. Counting unpublished papers we have read or seen mention of, there are probably close to two dozen contributions to the new urban economics.

2. Characteristics of the new urban economics models

■ Although new urban economics models differ from each other in important aspects, they characterize the city, or urban area, somewhat as follows. The city is built on a flat plain and travel is assumed to be equally costly in all directions. The city has a well-defined and predetermined central business district (CBD), and may have a pie slice of land removed from it to allow for natural features such as harbors or mountains. The CBD is usually assumed to be of fixed size and to employ a fixed number of workers, most typically the city's entire labor force. Most models do not describe the employment sector in further detail, although some recent papers are notable exceptions. For example, Mirrlees and Dixit introduce economies of scale in the production sector, and Solow introduces diffused local employment in addition to that in the CBD.³

In virtually all models, the only travel is commuting trips of the labor force between places of residence and work places in the CBD. Travel within the CBD is usually ignored. Thus, the only spatial characteristic of any location in the city that matters is distance from the CBD or, equivalently, from the city center. This is very important because it means that the residential area of the city can be treated as if it were one-dimensional. A one-dimensional representation of location seems to be crucial if the calculus of variations or control theory is to be used. Distance from the city center plays the role of time in more conventional applications of these techniques.

Travel either costs money or reduces utility (but not both so far, apparently for technical reasons). Thus, other things equal, people would like to live as close as possible to the CBD. But travel is of course not the only thing about which people have preferences. In earlier models, such as those of Muth and Mills,⁴ it was assumed that everyone's demand for housing could be described by a particular function of income and the price of housing. More recent models, such as those of Solow and Dixit,⁵ postulate utility functions defined

³ In [11], [4], and [19], respectively.

⁴ In [13] and [8], respectively.

⁵ In [17] and [4], respectively.

on consumer goods and housing. The latter approach is theoretically preferable, although the former has the advantages that the models are easier to work with and that the parameters of the demand for housing function can be verified empirically. There is nothing to prevent land from entering the welfare function directly, as in Mirrlees,⁶ but in most models it enters only indirectly via a housing production function which transforms goods and land into housing services.

In most models, the population of the entire urban area is given, and everyone is assumed to have the same utility or demand functions. People are allowed to live anywhere outside the CBD, and since they can travel in any direction at the same cost, the residential section of the city is doughnut-shaped. The size of the city is determined by incomes, tastes, housing technology, the cost and/or speed of travel, and how much must be paid to bid land away from nonurban uses such as agriculture.

The principal characteristics of most theoretical urban models are those described above. Particular models often have additional features, of course. For example, Muth⁷ and others have examined what happens when there are two or more groups of people with different incomes but the same utility functions. And a number of recent models, such as those of Dixit, Mills, and Solow,⁸ allow the speed or cost of travel to vary with traffic congestion and the amount of land devoted to transportation to vary with distance from the CBD.

So far, all these models have been static. It is too easy to criticize static models, and we do not intend to do so here. The problem is that theoretically satisfactory dynamic models seem to be enormously difficult to formulate and solve. Even for nonspatial systems economists know very little about dynamic general disequilibrium systems; and it is hardly fair to expect urban economics, where the spatial aspect makes everything so much harder, to progress faster than general economic theory. For most problems that are theoretically interesting, the choice seems to be between static models and no models at all. Those who practice the new urban economics have chosen the former.

Urban models can be solved in two quite different ways. Some models are normative, and for them a solution is an allocation of people, goods, housing, and land at each distance from the CBD which maximizes a social welfare function. Other models are positive, and for them a solution is a competitive equilibrium. The normative models are mostly solved by variational methods or control theory, whereas the positive models tend to be solved by ad hoc methods. In both cases explicit solutions are often difficult or impossible to find, and numerical analysis often has to be resorted to.

■ What has the new urban economics taught us? The most common focus has been on equilibrium or optimum population or housing density as a function of distance from the CBD. Remarkably simple models produce population densities that decline with distance from

3. Implications

⁶ See [11].

⁷ In [13].

⁸ In [4], [9], and [17], respectively.

the CBD in ways roughly consistent with data from real cities. In some models, it can be established that an equilibrium density pattern is optimum. Under some restrictive conditions it can be shown that, if the population consists of two or more groups which differ only in exogenously determined incomes, then the higher income groups live further from the CBD and in lower density neighborhoods than the lower income groups, in both equilibrium and optimum solutions.⁹

An especially interesting aspect of the density issue concerns congestion in the transportation system. The conditions that make it optimum to use land near the CBD intensively for housing also make it optimum to use close-in land intensively for transportation. But an important way to use land intensively for transportation is by congestion. It is clearly optimum to have more transportation congestion close to the CBD than in the suburbs, and some interesting calculations have been made of the way that optimum congestion might vary with distance from the CBD. The new urban economics supports the presumption that congestion will be different from the optimum in equilibrium, but little is known about the details of possible comparisons.

Recent urban models have also shed important light on the process of urban decentralization or suburbanization. Many urban specialists attribute suburbanization to racial conflicts or to the desire by middle and upper income groups to avoid high central city taxes. Models that exclude these factors cannot of course prove that they are unimportant. But recent urban models conclude without exception that decentralization will certainly result from increases in income and decreases in the time and cost of commuting. Undoubtedly, models that located employment endogenously within the urban area would provide an even fuller explanation of suburbanization. In fact, the evidence is that decentralization pervades cities throughout the developed world. The surprising thing is that anyone would have attributed it mainly to the parochial problems of U. S. cities.

Perhaps the most intriguing result of the new urban economics is the finding that an urban structure that maximizes social welfare may require that people resident at different distances from the CBD achieve different utility levels. This may be so even in a model in which all citizens have the same utility functions, endowments, and skills, and in which the social welfare function is penalized by inequality. Nevertheless, social welfare maximization may require that per capita utility increase or decrease with distance from the CBD. The basic results appear in papers by Mirrlees, Dixit, and Riley.¹⁰

Perhaps the most important issue that has been attacked by the new urban economics is the ability of competitive markets to sustain an optimum urban structure. Some models have demonstrated circumstances under which the equilibrium urban structure is efficient. More important, some have studied market failure and public policies to correct resource misallocation. Most attention has been paid to market failure because of externalities resulting from high residential density or congestion in transportation. It is easy to un-

⁹ See, among others, Muth [13], and Montesano [12].

¹⁰ See [11], [4], and [15], respectively.

derestimate the importance of this research unless one reads some of the literature that is written on urban problems without benefit of familiarity with welfare economics. City planning, the profession whose job is to improve on market resource allocation in urban areas, has been practically unaffected by welfare economics.

■ Where will the new urban economics go in coming years? There are two fairly distinct reasons for studying urban models. The more obvious reason is to prepare for realistic models of urban growth and structure. It is often easier and cheaper to formulate, manipulate, and test relationships in models that are deliberately simplified and do not require much empirical input. The analogy in macroeconomics is that an important justification for much of macro theory is the construction of better Brookings models. At present the only available urban model in this category that is the work of economists is the NBER model.¹¹ The second reason for building theoretical urban models is to gain insight into, or to give an account of, a basic characteristic of urban areas by a model that contains a small number of relationships thought to be crucial to the characteristic in question. Examples of basic characteristics are the relative locations of different income groups, centrality, variable capital-land ratios, congestion, and decentralization. The new urban economics models mostly fall in this category. One analogy in macroeconomics is, of course, the outpouring of neoclassical growth models in the 1960s.

It is certainly desirable and inevitable for more models of the second kind to appear in coming years. There are several issues that can and should be studied in such models. Further work is needed to clarify the issue of optimum unequal utilities in urban areas. Does the finding of optimum inequality carry over to more complex models? If so, what can be said about the optimum amount and spatial distribution of inequality? It would be desirable to introduce population types that are diverse in skills or endowments of productive assets. What can be said about the optimum residential mixing of such diverse groups? It should also be possible to introduce diversity of tastes, e.g., for neighbors with similar race, religion, or income. Clearly, more attention is needed to the production side of urban structure. If there are two or more production sectors, what is their optimum spatial distribution, and can it be generated by competitive markets? Further study is also needed of the transportation sector. It should be possible to establish general conclusions about the urban size and structure that justify particular modal mixes.

Our hesitation about the future of the new urban economics stems from the fear that large amounts of technical expertise might be expended on relatively minor variations on a few themes. Especially with optimization models, working out even relatively minor variations can be a considerable technical chore. The analogy with growth theory is irresistible. Many economists feel that some of the neoclassical growth theory literature in the 1960s consisted of technical pyrotechnics that added neither insight nor realism to previous work. In fact, for almost every theoretical growth model there is an analogous urban model in which distance is substituted

4. Future directions

¹¹ See [5].

for time. We are concerned that the profession will quickly see that Pontryagin's principle is as applicable to space as to time and that the journals will be flooded with technically sophisticated but economically uninteresting urban models. It is our judgment that much of the profession's effort at urban modeling should be devoted to making the models more realistic so that they can be used to plan public services, to evaluate private market performance, and to evaluate the desirability and efficacy of public regulation and control of private markets in urban areas.

We believe the profession faces an important issue in the method of formulating and analyzing more realistic models of urban structure. The rest of these notes present our views on this issue. The choice is between models in which space is treated as a continuous variable and models in which space is divided into a number of discrete sections. For convenience we refer to these as continuous and discrete models.

Continuous models are usually solved using variational methods or Pontryagin's principle. The advantage of these techniques is that, at least in fairly simple models, qualitative properties of the model can be derived. But in more complicated models a great deal of ad hoc numerical analysis is often required. A continuous model can be changed to a discrete model by dividing the land in the city into a finite number of sections, perhaps squares or circumferential rings. If the model contains n sections of land and m spatially undifferentiated goods, then the model is a general equilibrium system with $n + m$ prices. There are at least two different ways to formulate and solve such models. We discuss the two ways briefly and then indicate why we believe that discrete models are likely to prove more fruitful than continuous ones.

Economists have recently begun to solve nonlinear general equilibrium systems by using a class of fixed point algorithms first discovered by Herbert Scarf. Several of these algorithms and their applications to nonspatial equilibrium systems are described by Scarf and Hansen,¹² and a more effective algorithm is introduced by Kuhn and MacKinnon.¹³ With a little ingenuity it is possible to solve discrete urban models by using these algorithms; some preliminary efforts to do so are discussed by MacKinnon.¹⁴ The simplest approach is to divide the land outside the CBD into perhaps eight or ten rings, with all land in each ring assumed to be the same distance from the city center. Solution of the model means finding the equilibrium price of land in each ring. Since the solution process is almost entirely numerical, functional forms do not have to be chosen with so much care as they do for similar continuous models, and the model can be modified extensively without changing the method of solution. Unfortunately, this approach is only applicable when the object is to find an equilibrium, not when it is to find an optimum.

A promising approach to discrete optimizing urban models is to treat them as linear or nonlinear programming problems. Models of this type have been formulated by Mills.¹⁵ Production is by standard programming technology in which input-output coefficients can vary

¹² In [16].

¹³ In [6].

¹⁴ In [7].

¹⁵ In [9, 10].

with building height and, perhaps, with scale. Transportation takes place on a system whose capacity can be determined endogenously, and at costs that depend on congestion. The objective function is to minimize the cost of meeting stipulated goals, which might involve export quantities, consumption levels, etc. The location of production is determined endogenously; if there is a CBD, it comes out of the model rather than out of exogenous imposition. A bonus of the programming solution is that optimum prices emerge from the dual problem.

Both types of discrete model seem to offer sizeable advantages over continuous ones. One advantage is ease of use. To set up and solve a discrete model the first time usually involves considerable computer programming. But subsequently the model can be changed and re-solved at very small cost. Extra constraints can be added, functional forms can be changed, and, in the linear programming models, the number and type of sectors can be altered very easily, if the original programming has been done with foresight. To build a different continuous model, it is usually necessary to start from scratch again, deriving in most cases a new set of differential equations which have to be solved again by ad hoc numerical methods.¹⁶ This is an important advantage, because investigating the effects of modifications is basic to much of the research in urban economics. How else can one study, for example, the effects of zoning regulations or racial segregation in housing markets in a general equilibrium context?

Using discrete models, researchers are less constrained than when using continuous ones to choose simple functional forms. For example, in his equilibrium model, Solow,¹⁷ like many others, uses a logarithmic utility function, which implies that price and income elasticities of demand for housing are unity. He conjectures that the logarithmic form is not very limiting. We have no reason to doubt his judgment, but it is evidently difficult to find out just what difference other functional forms make. With the fixed point technique it should be possible to vary price and income elasticities routinely to determine their effects on the solution of the model. This is an important issue because the magnitudes of price and income elasticities of housing demand are major objects of current empirical research, and it is important to be able to incorporate the results of such research into urban models.

Similar remarks apply to other functions in urban models, such as housing production functions and functions representing transportation costs and technology. Discrete approaches permit the representation of a wide range of possible housing production technologies, whereas continuous approaches usually require that a simple production function, such as the Cobb-Douglas, be used. Likewise, continuous models require extremely simple transportation systems. In most, transportation costs can be monetary or psychic, but not both, and congestion cost functions are invariably extremely simple functions of the amount of congestion. Discrete models permit travel both to enter utility functions and to cost money. Speeds and costs may be varied from section to section and

¹⁶ See, for example, Solow [17, 19].

¹⁷ In [17, 18].

in almost any way with congestion, and people can be permitted to choose among different modes of transportation.

It is somewhat ironic that economists are becoming proficient at building models in which all employment is in the CBD and all housing outside it, when cities look less and less like that paradigm. Major metropolitan areas have numerous nuclei of employment and commercial activity, not just a single CBD. Understanding the density and location of employment is at least as important as modeling the residential sector. Continuous models have so far had little to say on the subject. Discrete equilibrium models can at least examine the effects on residential location of multiple employment centers. And linear programming models make the location of productive activity completely endogenous.

If there is more than one employment nucleus, location cannot be indexed just by distance from the CBD. Space must be represented in two dimensions. This dramatically increases the complexity of continuous models. Equilibrium models would require the solution of systems of partial differential equations, and the mathematics of control theory has not even been worked out for this problem. For discrete problems, on the contrary, two dimensional space raises no problems of principle. If large cities or many housing or production sectors are involved, discrete models may, however, run into severe computational problems.

The chief disadvantage of discrete models is, of course, their almost complete dependence on numerical analysis and consequent lack of qualitative results. Moreover, a discrete theoretical city is only an approximation to a "true" continuous one, and this tends to offend one's theoretical sensibilities. No doubt an optimum research strategy would entail substantial efforts with both continuous and discrete approaches. There is much to be learned from both. The experience with growth theory in the 1960s leads us to suspect that the profession is not likely to underestimate the benefits from building continuous urban models. We fear that they may underestimate the benefits from building discrete ones.

References

1. BECKMANN, M. J. "On the Distribution of Urban Rent and Residential Density." *Journal of Economic Theory*, Vol. 1, No. 1, 1969, pp. 60-67.
2. ———. "Von Thünen's Model Revisited: A Neoclassical Land Use Model." *Swedish Journal of Economics*, Vol. 74, No. 1 (March 1972), pp. 1-7.
3. BROWN, H. J. ET AL. *Empirical Models of Urban Land Use: Suggestions on Research Objectives and Organization*. New York: Columbia University Press for National Bureau of Economic Research, 1972.
4. DIXIT, A. "The Optimum Factory Town." *The Bell Journal of Economics and Management Science*, Vol. 4, No. 2 (Autumn 1973), pp. 637-651.
5. INGRAM, G. ET AL. *The Detroit Prototype of the NBER Urban Simulation Model*. New York: National Bureau of Economic Research, 1972.
6. KUHN, H. W. AND MACKINNON, J. "The Sandwich Method for Finding Fixed Points," forthcoming.
7. MACKINNON, J. "Urban General Equilibrium Models and Simplicial Search Algorithms," unpublished manuscript.
8. MILLS, E. S. "An Aggregative Model of Resource Allocation in a Metropolitan Area." *The American Economic Review*, Vol. 57, No. 2 (May 1967), pp. 197-210.
9. ———. "Markets and Efficient Resource Allocation in Urban Areas." *Swedish Journal of Economics*, Vol. 74, No. 1 (March 1972), pp. 100-117.

10. ———. "Mathematical Models for Urban Planning," forthcoming.
11. MIRRELES, J. "The Optimum Town." *Swedish Journal of Economics*, Vol. 74, No. 1 (March 1972), pp. 114–135.
12. MONTESANO, A. "A Restatement of Beckmann's Model on the Distribution of Urban Rent and Residential Density." *Journal of Economic Theory*, Vol. 4, No. 2 (April 1972), pp. 329–354.
13. MUTH, R. *Cities and Housing*. Chicago: Univ. of Chicago Press, 1969.
14. ORON, Y., PINES, D., AND SHESHINSKI, E. "Optimum vs. Equilibrium Land Use Pattern and Congestion Toll." *The Bell Journal of Economics and Management Science*, Vol. 4, No. 2 (Autumn 1973), pp. 619–636.
15. RILEY, J. G. "Gammaville: An Optimal Town." *Journal of Economic Theory*, forthcoming 1973.
16. SCARF, H., WITH THE COLLABORATION OF T. HANSEN. *The Computation of Economic Equilibria*. Forthcoming, Yale Univ. Press.
17. SOLOW, R. "Congestion Cost and the Use of Land for Streets." *The Bell Journal of Economics and Management Science*, Vol. 4, No. 2 (Autumn 1973), pp. 602–618.
18. ———. "Congestion, Density and the Use of Land in Transportation." *Swedish Journal of Economics*, Vol. 74, No. 1 (March 1972), pp. 161–173.
19. ———. "On Equilibrium Models of Urban Location," in J. M. Parkin, ed., *Essays in Modern Economics*, London: Longmans, 1973.
20. ——— AND VICKREY, W. "Land Use in a Long Narrow City." *Journal of Economic Theory*, Vol. 3, No. 4 (December 1971), pp. 430–447.

Congestion cost and the use of land for streets

Robert M. Solow

Professor of Economics
Massachusetts Institute of Technology

The traditional model of the equilibrium economic geography of a monocentric city is generalized to introduce congestion costs as well as distance costs of transportation. This has the effect of generating more curvature in the equilibrium rent gradient than the earlier theory had suggested. More important, if congestion tolls are not charged on crowded roads, land rents will reflect the private costs of transportation, not the full social cost. Market land values will then be faulty guides to land use, and cost-benefit analyses will give misleading results. Some sample computations suggest that the result will often be to cause more land to be taken for use as streets than is socially desirable, especially nearer the center of the city.

1. Introduction

■ In the standard equilibrium model of urban location, land values are determined primarily by differential accessibility to the Central Business District (CBD). Rents and land values are lower nearer the outer edge of the city. (Most such theoretical cities are monocentric and circular; in a more realistic model, with several centers of employment and perhaps alternative shopping and recreation districts as well, the equilibrium pattern of land values would be much more complicated, but the general principle would be the same.) Barring “distortions,” one might hope that competitive equilibrium land values would be a fair guide to the social value of land in different locations. If that were so, the market values of parcels of land could be used in cost-benefit analyses and land-use planning in a straightforward way.

But there is one distortion that is almost certainly present in all cities. It is well known that when streets or other transportation facilities are congested—in the sense that an additional trip by any traveler imposes delay costs on other travelers—they will be inefficiently used, unless a corrective toll is charged, which is hardly ever the case. The private cost of a trip, in time and expense to the traveler, will be less than the social cost of the trip, which includes the time costs imposed on others. These congestion costs are usually omitted from the standard models of urban location, which make the cost of a trip depend only on the distance covered—and perhaps the

Robert M. Solow received the A.B., the M.A., and the Ph.D. from Harvard University (1947, 1949, and 1951, respectively). He is Professor of Economics at M.I.T., where his research is in theoretical urban economics and the economics of exhaustible resources.

I want to thank Dr. John Broome of Birkbeck College, London, for planning and programming the numerical solutions, and Mr. Benjamin Wurzberger of M.I.T. for further computational assistance. I am grateful also to the National Science Foundation for support of this research.

particular route—but not on the volume of travel. As soon as congestion occurs, there is a mutual dependence: private travel cost depends on the volume of traffic, but the volume of traffic depends on the cost.

In any case, if land values reflect differential travel costs, they reflect private costs only. They will therefore give a distorted picture of the relative social value of parcels of land in different locations. In particular, one would expect the market to put too low a value on centrally located land relative to peripheral land, because the owner of centrally located land can capitalize only the private cost savings from close-in location, not the full social cost savings, and generally one would expect congestion to be greatest near the center. Cost-benefit analysis using the market value of land would then give wrong results.

This general insight provides the motivation for this paper. In an earlier article,¹ I extended the standard model of urban geography to include congestion costs of transportation. In that paper, however, I worked entirely with a single special case, using convenient but unrealistic parametric assumptions, simply because they permitted pencil-and-paper analysis. I listed a series of generalizations that would have to be made if the model were to supply any insight into the problem of urban land use. In this article, most of that program is carried out. As so often happens, the price is that numerical computation has to replace analytical solution. Most of the results that follow come in the form of sample computations, although I begin by describing the underlying theory, so that the paper is self-contained.

I hope that this article accomplishes three things. The first is to demonstrate that the presence of congestion cost should be expected to change rent gradients in a definite way—to make them more convex than standard theory has suggested. The second is to advance the presumption that when congestion is important, and corrective road tolls are not charged, the use of market values for land in costing transportation improvements will usually lead to too much land's being taken for streets, especially near the center of the city. The third is to suggest that large-scale simulation models of urban activity might usefully try to build in congestion phenomena.

The next section describes the model at the level of generality so far achieved. I then report some sample computations of rent and density gradients, to show how the equilibrium patterns for this model differ from those of its simpler predecessors. The next two sections take up the problem that prompted the investigation in the first place—the optimal use of urban land for streets. Finally, I suggest some further maneuvers that might repay analytical and computational effort.

■ I shall be brief here because an earlier article² contains more detail, and the general framework is traditional. The abstract city of the model consists of a circular Central Business District, of radius one,³ surrounded by an annular residential zone, whose

2. A model of residential geography with congestion

¹ In [8].

² See Solow [8].

³ In another paper [9] I show how the boundary of the CBD could be determined inside the model.

outer radius R is either given or determined by the condition that the urban residential rent at the boundary should just equal the given "agricultural" rent r_A . The city is inhabited by a given number N of families, each with one worker who works in the CBD. All travel is radial travel to and from the CBD. The N families are identical: they each have income y ; they each choose a location (i.e., distance from the CBD) x , an amount of space s , and consumption of a composite commodity in amount c , to maximize a utility function $U(c, s) = k \log s + (1 - k) \log c$, subject to the budget constraint $c + r(x)s + t(x) = y$. Here $r(x)$ is the rent per unit area of space at distance x from the CBD and $t(x)$ is the annual cost of transportation for a family living at x ; c is numeraire and its price is normalized at unity. In equilibrium, all families will have to achieve identical utility, no matter where they choose to live (that is, they must be indifferent among locations). Therefore $r(x)$ will have to be lower at greater distances, to offset the higher transportation costs. Those who live further away from the CBD and achieve the common level of utility will be induced by this lower rent to substitute space for the composite commodity; so $s(x)$ will be an increasing function of x (and its reciprocal, density, a decreasing function).

Let $n(x)dx$ be the number of families living in the ring with inner radius x and outer radius $x + dx$, and let $N(x) = \int_x^R n(u)du$ be the number of families living outside the circle of radius x . Suppose a fraction $b(x)$ of the land at x is used for residential purposes and the remaining $1 - b(x)$ is paved into roads. Then the aggregate "width" of the road network at x is $2\pi x[1 - b(x)]$, because the total land area in a ring of radius x and thickness dx is $2\pi x dx$. Now suppose that the annual cost of moving one person one unit of distance is t_0 plus a congestion cost component, and the congestion cost at x is proportional to the density of traffic at x , defined as the ratio of the number of commuters to the width of the road. I assume for simplicity that travel cost within the CBD is uniform for everyone so there is no harm in setting it at zero. Then $t(x)$, the annual travel cost for a commuter living at x , is obtained by adding up the cost per person per mile for each mile traversed between x and the boundary of the CBD:

$$t(x) = \int_1^x \left[t_0 + a \frac{N(u)}{2\pi u[1 - b(u)]} \right] du.$$

Conventional consumer theory tells us that each family spends on dwelling space a fraction k of income after transportation cost. Thus $r(x)s = k[y - t(x)]$. The condition that the decreasing rent gradient just offset the increasing travel-cost gradient is easily seen to be that $sr'(x) + t'(x) = 0$. From these two equations it follows that

$$r(x) = r_1 \left(1 - \frac{t(x)}{y} \right)^{1/k} = r_1 w(x)^{1/k}$$

and

$$s(x) = \frac{ky}{r_1} w(x)^{1-1/k},$$

where now r_1 is the residential rent at the edge of the CBD and has to be determined as a kind of constant of integration, and

$w(x) = 1 - \frac{t(x)}{y}$ is the fraction of income available after travel cost to a resident at x .

If $t(x)$ had no congestion-cost component, it would be a given function and these last two equations would state the equilibrium rent and density patterns explicitly, and that would be the end of the story. Congestion introduces a mutual dependence between transportation cost and the distribution of population and so $t(x)$ or $w(x)$ is not given and has to be solved for. To that end, it is better to transform the equation for $s(x)$ into one for density, or actually for $N(x)$, explicitly. From the definition, $N'(x) = -n(x)$. Now $n(x)dx$ is the number of families living in a thin ring at distance x , and each of them consumes an amount of space $s(x)$. The area of the ring is $2\pi x dx$, and $2\pi x b(x) dx$ is available as residential space. In equilibrium, if every family's demand for space is to be satisfied, it must be that $s(x)n(x)dx = 2\pi x b(x)dx$. It follows that

$$N'(x) = -\frac{2\pi r_1}{ky} b(x)xw(x)^{1/k-1} = -m_1 b(x)xw(x)^{1/k-1}. \quad (1)$$

Next, differentiate the equation for $t(x)$ and substitute $t'(x) = -yw'(x)$. The result is

$$w'(x) = -\frac{t_0}{y} - \frac{a}{2\pi y x[1-b(x)]} \frac{N(x)}{x} = -\frac{t_0}{y} - m_2 \frac{N(x)}{x[1-b(x)]}. \quad (2)$$

Equations (1) and (2) are a pair of first-order differential equations for $N(x)$ and $w(x)$; the first of them describes the dependence of equilibrium population density on travel cost, and the second shows how travel cost depends on population density. (If a were zero, so all travel cost were distance cost, (2) would degenerate.) In these equations, t_0 , k , a , and y are known constants and $b(x)$ is a known function for x between one and R , taking values between zero and one: it is part of a city-planning decision. The number r_1 is still to be determined.

Even if r_1 were given, these equations would need two boundary conditions to be fully determinate. Since r_1 is not given, they need three. They are:

$$\begin{aligned} N(1) &= N \\ N(R) &= 0 \\ w(1) &= 1. \end{aligned} \quad (3)$$

The last comes from the convention that travel cost inside the CBD is zero, and would have to be fixed up if a more realistic assumption were made. There is now sufficient information to determine the equilibrium $N(x)$, $w(x)$, or $t(x)$, and then we have formulas for $s(x)$ and $r(x)$. (If r_A were specified instead of R , the solution could proceed by trying different values of R until one was found for which $r(R) = r_A$.)

For some purposes one would want to transform these two first-order equations into a single second-order equation. This is easily

done:

$$w''(x) + w'(x) \left[\frac{1}{x} - \frac{b'(x)}{1 - b(x)} \right] - m_1 m_2 \frac{b(x)}{1 - b(x)} w(x)^{1/k-1} = -\frac{t_0}{y} \left[\frac{1}{x} - \frac{b'(x)}{1 - b(x)} \right], \quad (4)$$

with boundary conditions

$$\begin{aligned} w(1) &= 1 \\ w'(1) &= -\frac{t_0}{y} - m_2 \frac{N}{[1 - b(1)]} \\ w'(R) &= -\frac{t_0}{y} \end{aligned} \quad (5)$$

and $r_1 w(R)^{1/k} = r_A$ if desired.

Nothing has been said so far about the ownership of land and the disposition of rent payments. In real life some people own land and collect and spend rent, others pay rent, and others are owner-occupiers. Since I intend to preserve for now the fiction that all families are identical, there are two polar land-tenure possibilities for the model: either nobody owns any land (i.e., all the land in the city is owned by absentee landlords who spend the rent elsewhere) or the land is owned by the city, which collects rent and distributes the proceeds as an equal lump-sum divided among the N families.⁴ In the first case, the income y that appears in the model is all wage income. In the second case, y is partly wage and partly the rent dividend.

If all that is at stake is a numerical example, one can solve the model for some arbitrary y , calculate the per family rent dividend in equilibrium, subtract it from y , and thus find the annual wage income for which this situation is a full equilibrium. If one knows the annual wage to start with, it is not hard to find a systematic iteration that starts with an arbitrary y_0 and converges to a y^* equal to the given wage plus the *equilibrium* rent dividend for income y^* .

This becomes important if one wishes to analyze, as I do, the welfare consequences of alternative decisions about $b(x)$, say, the allocation of urban land between streets and dwellings. A change in $b(x)$ will change the aggregate rent payments. If the land is owned by absentees, an increase in aggregate rents is a dead loss to the inhabitants of the city. If the land is owned by the city itself, the aggregate volume of rent payments is *per se* unimportant, since the inhabitants owe it to themselves. Of course the change in $b(x)$ has real consequences, because it either reduces the amount of dwelling space available and diminishes aggregate travel costs, or increases both. We will come to that later.

3. Choice of parameters

■ When $k = \frac{1}{2}$ —and only then—(1) and (2) form a system of linear differential equations. That was the special case I treated in the earlier paper,⁵ when it turns out that the solution is in terms of

⁴ I am grateful to Dr. D. Pines of Tel-Aviv University for pointing out to me that, in the earlier paper, I write as if the public-ownership case were valid but calculate as if the absentee-landlord case held.

⁵ See [8].

standard tabulated functions. But $k = \frac{1}{2}$ means that each family spends half its income after travel costs on dwelling space and the other half on other consumption, which hardly ranks as a first approximation to the facts. As soon as k is less than $\frac{1}{2}$, (1)–(2) is a non-linear system. Even then, it is possible to learn something about the qualitative character of solutions by pencil-and-paper analysis.⁶ In the end, however, one has to go over to numerical solution of the differential equations for specific choices of the parameters and the function $b(x)$.

In all cases, I have taken $N = 300,000$. The abstract city has that many families and thus about a million people. The radius of CBD is always one. When the outer radius of the city is given, I have taken $R = 4$. For these choices to make approximate sense, the unit of distance should be about two miles (which I shall call a “league,” just to have a word for it). The total area of the city is thus about 200 square miles (or 50 square leagues) of which a bit over 12 square miles is CBD and the rest residential.

Wage income is \$10,000 for each family. When aggregate land rents are collected by the city and redistributed equally, total family income is higher than that, usually around \$13,000.⁷ Remember that “consumption” is numeraire.

The choice of a value for k presents a problem. Land accounts only for a small fraction of housing costs; Muth⁸ suggests about 5 percent, or a bit more, depending on how one treats certain improvements. If one interprets k strictly as the fraction of income after travel cost spent on ground rent, then k should be a very small number, perhaps 0.02. It seemed to me, however, that this might give a distorted picture because the model, as it now stands does not distinguish housing as a separate commodity. Social custom suggests a strong complementarity between land and house, so that anyone who buys a large and expensive lot is more or less committed to a large and expensive house. For some computations, I have set $k = \frac{1}{4}$ as if house and lot were perfectly complementary. In these calculations, rent stands for the annual rent of land-plus-house. This is not a satisfactory resolution, because the complementarity between house and lot is certainly not perfect and consumers can and do substitute one for the other. For this reason I did a few solutions with $k = \frac{1}{10}$, just to see what difference it makes. (Naturally redistributed rents then amount to only about \$1000 per family.)

A better way out of the dilemma would be to include housing explicitly as a commodity, and that is definitely on the agenda. One could register both house and land as expenditure items and as separate variables in the consumer’s utility function. A slightly less general, but in some ways more attractive, idea is to identify a single variable in the utility function called “housing services.” Housing services are produced according to a given technology from inputs of land and intermediate commodities in variable proportions. The price of

⁶ I did not originally see how. But I was shown a few tricks by my colleague Professor Frederic Wan of the M.I.T. Mathematics Department. For $t_0 = 0$ and $b(x) = \text{constant}$, a few transformations lead to an enlightening phase diagram.

⁷ That is when $k = \frac{1}{4}$. The exact aggregate relation among rent, income, and transport cost will be discussed and used later.

⁸ In [6], p. 53.

housing services is generated by the rent of land and the price of commodities. In that formulation, k would apply to the composite housing services and would indeed be about $\frac{1}{4}$. But minimum-cost production of housing services would permit housing expenditures to be decomposed determinately into a ground rent and a house rent. This is the way I intend to proceed, but I have nothing to report now.

That leaves the land-use function $b(x)$ and the transportation-cost parameters t_0 and a . It will be remembered that $b(x)$ is the fraction of land area at distance x from the center that is devoted to housing, with $1 - b(x)$ the fraction in streets. Obviously the simplest choice is to make $b(x)$ constant; a glance at (4) will show how this simplifies the differential equation. A reasonable value for b is then somewhat around 0.8.⁹ But, of course, real cities do not have $b(x)$ constant. They devote more land to streets near the center and less near the periphery, and for good reason. The central portions of the road network carry more traffic than the outer portions even in our simple radial pattern, because only those who live outside x use any part of the road system outside of x . If $b(x)$ were constant, there would be less road area near the center [because the road area at x is $2\pi(1 - b)x$] carrying more traffic, hence at higher density, hence with much more congestion. It would clearly be sensible to widen the roads near the center even at the expense of narrowing them in the lightly travelled outer portions. That is to say, it would be better if $b(x)$ were an increasing function. I will discuss this sort of question in more detail later on, when the problem is actually to find a best $b(x)$. In the meanwhile, I shall report some calculations in which $b(x)$ is linear and increasing in x . In choosing values for t_0 and a , the general idea has been to generate equilibrium patterns in which the median family (in terms of distance from the center) spends about 10 percent of its income on travel, because that appears to be about the right number. Subject to this constraint, one can try alternative pairs; a high value of a and correspondingly low value of t_0 represent a situation in which pure congestion costs are high relative to pure distance costs, and *vice versa*.¹⁰

There is a conceptual problem here, too. Cash outlays on transportation do indeed run typically to about 10 percent of income. But cash expense is only part of the cost of travel to the commuter; the other part is the use of the commuter's own time. Most studies seem to suggest that travelers value their travel time at something near half their wage. If one were to try to account for this cost explicitly, I suppose the time costs ought to be added both to income and to travel costs. In any case, I have not tried to deal explicitly with the time costs of travel. I have experimented with higher values of both a and t_0 to generate travel costs in excess of a tenth of income, though I am not sure how much the excess ought to be. Perhaps a full dress treatment of the allocation of time would be a worthwhile extension if the basic model proves to be interesting and enlightening, but that is for later work.

⁹ New York has 30 percent of its area in streets, Los Angeles 14 percent, and Phoenix 10 percent. See [3], p. 4.

¹⁰ It will be more convenient to return later to the proportional division of travel costs between distance and congestion corresponding to any particular pair of values for a and t_0 .

■ The object of this section is merely to exhibit a few numerical solutions of the differential equations (1) and (2), with boundary conditions (3). The idea is to show how the rent and density gradients vary with the parameters of the model. In particular, when $a = 0$, so that all transport costs are distance costs, this model reduces to the standard model of Alonso, Muth, Mills and others.¹¹ So the influence of congestion on rent and density can be isolated by comparing solutions in which a is very small or zero with solutions in which a is relatively large but t_0 is small, so long as commuting costs eat up roughly the same fraction of income in both cases.

TABLE 1
SAMPLE EQUILIBRIUM SOLUTIONS

DISTANCE x	RENT $r(x)/1000$		SPACE PER FAMILY $s(x) \cdot 100,000$	
	A	B	A	B
1.000	54,715	54,636	5.84	5.83
1.500	45,143	42,920	6.75	6.99
2.000	36,885	34,477	7.85	8.23
2.500	29,816	28,057	9.21	9.61
3.000	23,815	23,126	10.90	11.11
3.369	20,019		12.42	
3.432		20,007		12.39
PARAMETERS:				
$N = 300,000, \gamma = 10,000, r_A = 20,000,000$				
$B_0 = 0.65, B_1 = 0.95, k = 1/4$				
CASE A: $t_0 = 1200, a = 0$				
CASE B: $t_0 = 700, a = 0.007$				

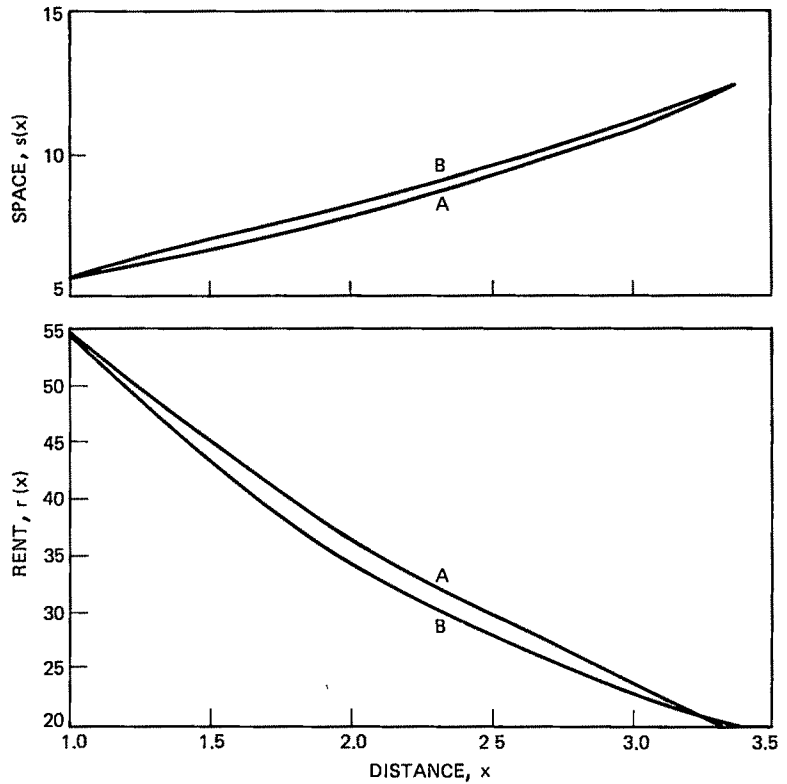
For example, Table 1 and Figure 1 compare two cases with $k = \frac{1}{4}$. The function $b(x)$ is linear, going from $B_0 = 0.65$ at the boundary of the CBD to $B_1 = 0.95$ at the outer perimeter of the city, so that the fraction of land in roads rises from 5 percent at the city limits to 35 percent at the edge of the CBD. The radius of the city was determined by setting the agricultural rent equal to 20 million (units of consumption per square league). The two solutions differ only in the transport-cost parameters: in Case A, $t_0 = 1200$, $a = 0$, while in Case B, $t_0 = 700$, $a = 0.0070$. Thus Case A illustrates a model without congestion, which is introduced in Case B. Aggregate commuting cost is almost identical in the two solutions. The parameters are chosen to make this aggregate bear the right proportion to total income.

The table and the diagram show the rent and the space occupied per family (the reciprocal of the density) as functions of distance from the center. The most obvious difference is the sharper curvature of the rent function in the presence of congestion costs: the two rent functions are equal at the inner and outer boundaries of the residential zone, but that for Case B lies about 7 percent below the one for Case A at a radius of two to three leagues. It is easy to understand why this should be so. Travel cost in Case B rises more sharply with distance near the CBD because of heavy congestion, and less sharply

4. Sample solutions of the model

¹¹ In [1], [6], and [4], respectively.

FIGURE 1
TWO EXAMPLES OF RENT AND SPACE FUNCTIONS



than in Case *A* near the periphery because congestion is negligible and t_0 smaller.

A second systematic difference is that the equilibrium area of the city is somewhat larger in Case *B*, again for obvious reasons. When distance costs predominate, contraction occurs because urban rent falls sharply with distance of the city limits; when congestion costs bulk larger, it pays to spread the population out, because distance from the center matters less and lower density reduces commuting cost.

It is an unsatisfactory aspect of these choices of parameters that the rent function falls very gently (and correspondingly the space function rises very gently) with distance. The rent at the edge of the CBD comes out as $2\frac{1}{2}$ times the rent at the periphery. And families nearest the center occupy about half as much space as the outermost suburbanites. Case *C* in Table 2 illustrates a solution when $k = \frac{1}{10}$. This solution is not otherwise directly comparable with *A* and *B* because the other parameters differ, so as to give sensible orders of magnitude. The agricultural rent is reduced to 1.5 million. The equilibrium city now extends almost five leagues from the center; the families on the periphery spend a quarter of their income on travel, only 2 to 3 percent more than the outermost families in Cases *A* and *B*. The main point about Case *C* is that rents at the inner radius are almost twenty times rents on the periphery, and the suburban family occupies ten times as much land as the inner-city dweller.

Case *D* is like Case *C* in that $k = \frac{1}{10}$, but the travel-cost parame-

TABLE 2
SAMPLE EQUILIBRIUM SOLUTIONS

DISTANCE x	RENT r (x)/1000		SPACE PER FAMILY s (x) • 100,000	
	C	D	C	D
1.00	25,372	72,798	4.32	1.51
1.50	15,646	32,840	6.67	3.08
2.00	10,341	16,476	9.69	5.73
2.50	7,096	8,620	13.59	10.27
3.00	4,982	4,558	18.69	18.22
3.50	3,552		25.34	
3.85		1,500		49.55
4.00	2,561		34.01	
4.87	1,500		55.04	
PARAMETERS: N = 300,000, $\gamma = 10,000$, $r_A = 1,500,000$, $k = 1/10$				
CASE C: $t_0 = 500$, $a = 0.0050$			$B_0 = 0.65$, $B_1 = 0.90$	
CASE D: $t_0 = 1000$, $a = 0.0100$			$B_0 = 0.50$, $B_1 = 0.70$	

ters are both higher (and more of the area is in roads). The result is that the city contracts to an outer radius of less than four leagues, and even so the outermost families spend a third of their income on travel. Rents at the inner radius are now almost fifty times those at the boundary; a family at the city limits occupies thirty-three times as much space as a family living at the edge of the CBD.

The last examples of this section, Cases *E*, *F*, and *G* in Table 3

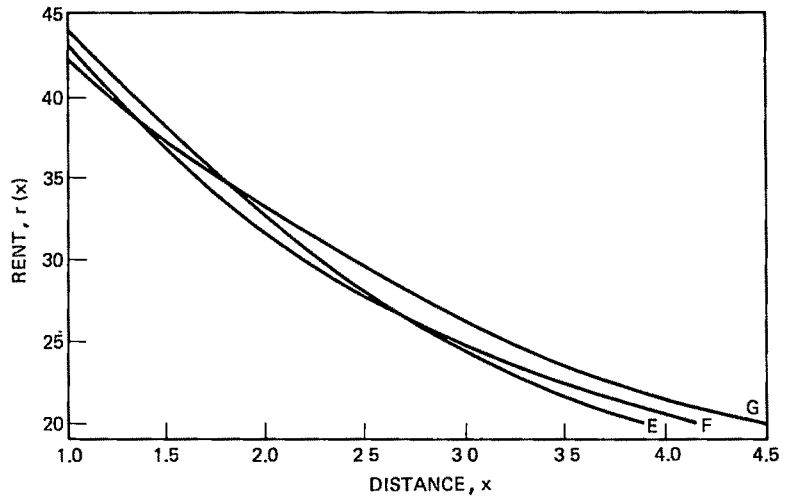
TABLE 3
SAMPLE EQUILIBRIUM SOLUTIONS

DISTANCE x	RENT r (x)/1000			SPACE PER FAMILY s (x) • 100,000		
	E	F	G	E	F	G
1.0	44,016	42,939	42,250	7.28	7.46	7.58
1.5	37,117	36,305	36,756	8.28	8.47	8.42
2.0	31,937	31,463	32,491	9.26	9.42	9.24
2.5	27,787	27,683	28,978	10.28	10.37	10.06
3.0	24,370	24,655	26,016	11.35	11.32	10.91
3.5		22,240	23,511		12.23	11.77
3.9	20,000			13.16		
4.0		20,406	21,439		13.04	12.62
4.1		20,000			13.24	
4.5			20,000			13.29
PARAMETERS. N = 300,000, $\gamma = 10,000$, $r_A = 20,000,000$						
$t_0 = 350$, $a = 0.0105$, $k = 1/4$						
CASE E: $B_0 = 0.41$, $B_1 = 0.95$						
CASE F: $B_0 = 0.41$, $B_1 = 0.82$						
CASE G: $B_0 = 0.20$, $B_1 = 0.82$						

and Figure 2 go back to $k = \frac{1}{4}$, and are intended to illustrate the effect of changing the amount of land devoted to roads. In Case *E*, the fraction of land in roads falls linearly from 59 percent at the edge of the CBD to 5 percent at the edge of the city; in Case *F* it falls from

FIGURE 2

EFFECT ON RENT FUNCTION OF DIFFERING AMOUNTS
OF LAND DEVOTED TO ROADS



59 percent to 18 percent, and in Case G from 80 percent to 18 percent. So the transition from Case E to Case F adds roads mainly near the periphery, and the transition from F to G adds roads mainly near the CBD. It will be seen that adding roads flattens the rent gradient, and increases the equilibrium diameter. The first result follows from any cheapening of transportation, since rent differences are travel-cost differences. The second occurs both for that reason and because there is less residential land left in any given area; because families have more to spend on rent and less land to spend it on, they bid land away from agriculture. Generally speaking, the flattening effect is more striking when roads are added near the CBD; this seems natural, because that is where most of the congestion is.

5. Aggregate relations between rent and transport cost

■ Some calculations will show how the measurement of some economic aggregates for the city can be simplified. Each family living in the ring at radius x spends on rent the fraction k of its total income net of travel cost. Its total income is its earnings, z , plus the social dividend D . The social dividend, the same for everyone, is calculated as aggregate rent payments divided by the number of families. (This “dividend” is an actual cash payment—the equal redistribution of rents collected by the municipality as landowner.) Therefore,

$$D = N^{-1} \int_1^R r(x)s(x)n(x)dx = N^{-1}k \int_1^R [z + D - t(x)]n(x)dx \\ = k(z + D - T),$$

where $T = N^{-1} \int_1^R t(x)n(x)dx$ is simply average travel cost per family. There is a simple linear relation connecting the average rent payment and the average travel cost:

$$T = z - \frac{1-k}{k} D. \quad (7)$$

Thus, for instance, $T = z - 3D$ if $k = \frac{1}{4}$, and $T = z - 9D$ if $k = \frac{1}{10}$.

For each numerical solution of the model, it is straightforward to compute an approximate value of T or D , in which case (7) gives the other.

Now consider the choice of a land-use pattern, restricted in this simple model to the allocation of land between road and residence. To begin with the simplest case, suppose that $b(x)$ must be constant and only the particular constant can be chosen. Then $D = D(b)$ and $T = T(b)$ are both functions of b , but of course they satisfy (7) for every b .

Imagine that a tentative value of b has been chosen, and it is desired to compare this choice with the alternative b' . One way to do so is to make a "cost-benefit analysis" of the change. The benefit of changing from b to b' is the saving in travel costs, i.e., $T(b) - T(b')$. This will be positive if b' is less than b , meaning that more land is devoted to roads. Using (7), the benefit per head turns out to be

$$T(b) - T(b') = \frac{1-k}{k} [D(b') - D(b)].$$

The only cost I shall account for is the opportunity cost of the land taken for road use in passing from b to b' , calculated at market value. The area of land taken is $(b - b')$ times the total area of the city; but initially only a fraction b of the city's area is residential. The change from b to b' thus takes a fraction $(b - b')/b$ of the residential land for conversion to road. Moreover, the change takes the same fraction of residential land at every distance x from the center.

I take it that the opportunity cost of land is measured as the differential rent, the excess of rental value over the rent at the city limits, r_A . Call opportunity cost per family $C(b)$. Then $C(b) = D(b) - \frac{\pi(R^2 - 1)r_A}{N}$. The "cost" of the change from b to b' is $\frac{(b - b')C(b)}{b}$ per family. Since $C(b') - C(b) = D(b') - D(b)$, the ratio of benefit to cost is

$$- \frac{1-k}{k} \frac{b[C(b) - C(b')]}{b - b'} \frac{1}{C(b)}.$$

If b and b' are very close together, this amounts to

$$- \frac{(1-k)}{k} \frac{bC'(b)}{C(b)} = - \frac{(1-k)}{k} e_C,$$

where e_C is the elasticity of opportunity cost with respect to b , the fraction of land in residential use.

No equally simple formula can be found for an arbitrary change in the land-use pattern, not even a small change, because the land cost will depend on the particular pattern of land-taking and on the whole rent function. But it is possible to generalize the formula given above to a comparison between any $b^1(x)$ and $b^2(x) = mb^1(x)$. It remains true that the social benefit is $T(b^1) - T(b^2)$

$= \frac{1-k}{k} [C(b^2) - C(b^1)]$. The market value of the land taken is¹²

$$2\pi N^{-1} \int_1^R x[b^1(x) - b^2(x)]r^1(x)dx,$$

where $r^1(x)$ is the opportunity-cost function in the initial situation. But then if $b^2 = mb^1$, the land cost reduces to

$$(1-m)2\pi N^{-1} \int_1^R xb^1(x)r^1(x)dx = (1-m)C(b^1).$$

The benefit-cost ratio is then

$$- \frac{1-k}{k} \frac{C(b^1) - C(b^2)}{C(b^1)} (1-m)^{-1}, \quad (8)$$

which reduces to the earlier formula if b^1 and b^2 are constant functions with $b^2 = mb^1$.

6. Optimal land use and cost-benefit analysis

■ Suppose that the only planning decision available to the municipality is the allocation of land between road and residence. There is no possibility—as there is in the papers of Mirrlees, Riley, and Dixit¹³—of assigning families to space and to consumer bundles, or of redistributing income through taxes and transfers. Then a policy is simply a choice of a $b(x)$ that makes the common utility level as large as it can be made by any choice.¹⁴

The first object of this section is to exhibit some optimal policies within this rather sparsely furnished model. The policies in question are not actually fully optimal; they are merely the best that could be found among *linear* $b(x)$. They were found by simple search over the parameters; nothing guarantees that there are not other local maxima, even better ones, but no symptoms of this multiplicity were ever discovered in numerical experimentation.

The second object is to see whether cost-benefit analysis of the simple kind already discussed is an adequate guide to the value of alternative land-use patterns. The question is worth asking precisely because congestion plays a role in the model. In the absence of congestion tolls, land values reflect differences in private travel cost, which is not the same as social cost. A family contemplating a move outward, compares the lower rent with the higher travel cost it must pay, but ignores the effect of its outward move in imposing congestion costs on a larger number of other families.¹⁵ The question is whether this distortion is substantial enough to make cost-benefit analysis seriously misleading, when it values land at market price.

Table 4 shows what was found to be the best linear land-use pattern for parameters $k = \frac{1}{4}$, $t_0 = 350$, $a = 0.0105$. To avoid the end-

¹² I am ignoring an end effect here, if the city expands or contracts as a result of the change in land use. In this calculation $m < 1$, so that land is being taken for roads. If $m > 1$, so that residential area is being enlarged at the expense of road area, then (8) is the ratio of cost to benefit.

¹³ In [5], [7], and [2], respectively.

¹⁴ It is shown in [8] that this common level of utility can be calculated apart from an additive constant, as $V = \log y - k \log r(1)$.

¹⁵ Namely those living in the ring between initial and final location.

TABLE 4

AGGREGATE TRANSPORT COSTS AND UTILITY
FOR DIFFERENT ROAD PATTERNS

B_0	B_1	D	T	V
0.4600	1.0000	2758	1726	5.0733
0.4554	0.9900	2765	1705	5.0730
0.4700	1.0000	2753	1741	5.0733
PARAMETERS: $N = 300,000$, $y = 10,000$, $r_A = 20,000,000$ $t_0 = 350$, $a = 0.0105$, $k = 1/4$.				

effects associated with shifts in the city boundary as land-use changes, the outer radius of the city was fixed at four leagues in this computation. The best linear $b(x)$ rises from $B_0 = 0.4600$ at the edge of the CBD to $B_1 = 1.0000$ at the edge of the city. In other words, the road network occupies 54 percent of the land at the inner radius of the residential zone and shrinks to zero as it approaches the city limits.¹⁶ For this choice, the common level of utility was $V = 5.0733$. Rent payments per family were 2758, so aggregate travel cost per family was $T = 10,000 - 3(2758) = 1726$.

Now I can compare this planning choice with another that adds another one percent of the land area at each x to the road network. This amounts to passing from $b(x)$ to $0.99b(x)$, so that (8) applies with $m = 0.99$. V falls to 5.0730, the dividend D rises to 2765, so T falls to 1705. To apply (8), I need $C(b) = 2758 - \frac{(16-1)\pi r_A}{300,000} = 2758 - 2525 = 233$. The ratio of benefit to cost is then calculated to be $3(7)(100)/233$ which is nearly 10. The distorted rent gradient causes the cost-benefit comparison to give the wrong result and instructs the road department to move away from the optimal land-use pattern in the direction of building too many roads. W. S. Vickrey and I have found a similar thing to happen in an urban-travel model of quite a different kind.¹⁷

The symmetric alternative of subtracting one percent of the land area from the road at each x is not available because the road shrinks to nothing at the city limit. I have, however, solved the model for $B_0 = 0.4700$ and $B_1 = 1.000$, just to see what happens. The average rent is 2753. The formula (8) does not apply directly. But one can calculate that T rises to 1741, 15 higher than in the optimal arrangement. The benefit corresponding to this cost is an increase in residential space of about 2 percent nearest the center, falling toward zero at the edge of the city. If one percent is a fair average of the increased residential space, then the social benefit should come to about one percent of 233, the opportunity cost, considerably less than the cost of 15. This move is rejected. Clearly cost-benefit analysis using market land values is strongly biased in the direction of building too many roads.

¹⁶ The fully optimal $b(x)$ must presumably go to unity at the city limit, because the volume of traffic shrinks to zero and it would be wasteful to have a road of finite width. There would not seem to be any mathematical need for the best linear $b(x)$ to do so, but in this case it does.

¹⁷ In [10].

The case with $k = \frac{1}{10}$ strains the capacity of the computational routine, so that I have not actually found a "best linear" land-use pattern. The accuracy of cost-benefit analysis using market land values can, however, be tested. For example, Table 5 records the

TABLE 5
AGGREGATE TRANSPORT COSTS AND UTILITY FOR
DIFFERENT ROAD PATTERNS

B_0	B_1	D	T	V
0.6370	0.8722	958.22	1376	7.5970
0.6435	0.8811	957.59	1382	7.5970
0.6500	0.8900	956.95	1387	7.5968
0.6565	0.8989	956.30	1393	7.5967
0.6630	0.9078	955.62	1399	7.5964
PARAMETERS: $N = 300,000$, $\gamma = 10,000$, $R = 4.87$ $t_0 = 500$, $a = 0.0050$, $k = 1/10$				

values of D , T , and V for an example with $k = \frac{1}{10}$, $t_0 = 500$, $a = 0.0050$, $B_0 = 0.6500$, and $B_1 = 0.8900$, and compares them with the solutions corresponding to uniform one-percent and two-percent additions to and subtractions from the residential area. In all cases, the outer radius of the city has been fixed at 4.87 leagues, so that minor end effects should not obscure the main point. (The rent at the city limits is then not the same in all the cases, but varies by 5 percent on either side of 1.5 million units of output per square league.)

It can be read from the table that each successive one-percent expansion of the residential area adds about six units of output per family to annual commuting cost. But average annual rent per family comes to about 957 units of output, and the standard computation shows that the opportunity cost of land is about 600 per family. So expansion of the residential area by one percent adds a net rental value of about six units of output per family more or less. To this degree of accuracy, therefore, the cost-benefit analysis comes out about even: the extra residential land is worth just about the increase in commuter cost generated by diminishing the road network. The table shows, however, that in this configuration, widening the road network increases the common achieved level of utility per family. Here, if anything, the conclusion of the previous example is reversed. Cost-benefit calculation using market values for land leads, if anything, to underuse of land for roads. But the calculation may simply not be sufficiently precise to give a distinct conclusion.

The final set of solutions, reported in Table 6, produces an unexpected result. Here, too, $k = \frac{1}{10}$, but the transportation-cost parameters, t_0 and a , are each twice as large as in the previous table. The outer radius of the city is fixed at 3.78 leagues. We start with $B_0 = 0.5$ and $B_1 = 0.8$, arbitrarily, and the table shows what happens when B_0 and B_1 are both increased and both decreased by one percent and then by 5 percent. The remarkable result is that as the residential zone widens (i.e., the road network narrows), average commuting cost per family *falls*. This is at least mildly counter-intuitive, because the narrowing of the road network ought to generate higher congestion costs. Apparently this effect is more than

TABLE 6

AGGREGATE TRANSPORT COSTS AND UTILITY FOR
DIFFERENT ROAD PATTERNS

B_0	B_1	D	T	V
0.4750	0.7600	954.74	1407	7.4917
0.4950	0.7920	954.59	1409	7.4939
0.5000	0.8000	954.39	1410	7.4944
0.5050	0.8080	954.51	1409	7.4949
0.5250	0.8400	954.61	1409	7.4965
PARAMETERS: $N = 300,000$, $\gamma = 10,000$, $R = 3.78$ $t_0 = 1000$, $a = 0.0100$, $k = 1/10$.				

balanced by a change in equilibrium residential density: as the road network narrows, families crowd in a bit closer to the center, enough so that the saving in average distance travelled outweighs the cost from more crowded roads. (I should, however, mention the possibility that this result is mere “experimental error.” In some of these sample solutions, the last few digits may not be significant; the numerical integrations are not all so accurate. They could be improved, of course, but I have not done that.) The Highway Department ought then to reduce the allocation of land to roads, if it takes this sort of analysis seriously. In this case, that is the correct thing to do: utility per family rises as the residential area expands.

■ The normal presumption in this kind of model is that calculations of costs and benefits using market values ought to lead to excessive road-building.¹⁸ The reason is this: there are no congestion tolls, so those commuters living away from the CBD do not bear the full social costs of their use of the roads. Thus the rent gradient is too flat; land values do not fall so fast as they ought, nor so fast as they would if corrective tolls were imposed. But the rent function is anchored at the outer edge of the city by the residual (agricultural or other) value of land. Therefore, rents nearer the center are too low to reflect properly the social worth of the transportation costs avoided by close-in location. In other words, the market rent function lies everywhere below the “correct” rent function. The value of land taken for roads is underestimated in cost-benefit calculations, and so too much of it will be taken. If this kind of model is worth following up, the first order of business probably ought to be to refine the computational scheme so that more precise numerical results can be obtained to estimate the magnitude of the distortion created by the absence of a road-pricing scheme.

As for structural improvement of the model, I would suppose the main items are: (1) the explicit inclusion of housing in addition to land as a residential cost; (2) the allowance for both time costs and out-of-pocket costs of commuting; (3) the use of a congestion-cost function that rises more than proportionally with traffic density; and (4) the existence of two or more income classes. If, as I am inclined to think, the main function of simple models like this one is to serve as preliminary exploration for larger simulation models, then it is

7. Final note

¹⁸ As in [10].

not necessary that all these improvements and extensions be made simultaneously.

References

1. ALONSO, W. *Location and Land Use*. Cambridge: Harvard Univ. Press, 1964.
2. DIXIT, A. K. "The Optimum Factory Town." *The Bell Journal of Economics and Management Science*, Vol. 4, No. 2 (Autumn 1973), pp. 637-651.
3. GOLDBERG, V. P. AND KRIESEL, K. M. "Urban Form and the Allocation of Land to Streets." Research Report No. 22 (duplicated), Institute of Governmental Affairs, University of California, Davis, October 1971.
4. MILLS, E. S. *Urban Economics*. Oakland, N.J.: Scott Foresman, 1972.
5. MIRRELES, J. A. "The Optimum Town." *Swedish Journal of Economics*, Vol. 74, No. 1 (March 1972), pp. 114-135.
6. MUTH, R. F. *Cities and Housing*. Chicago: Univ. of Chicago Press, 1968.
7. RILEY, J. G. "Gammaville: an Optimal Town." Forthcoming in *Journal of Economic Theory*, 1973.
8. SOLOW, R. M. "Congestion, Density and the Use of Land in Transportation." *Swedish Journal of Economics*, Vol. 74, No. 1 (March 1972), pp. 161-173.
9. ———. "On Equilibrium Models of Urban Location," in J. M. Parkin, ed., *Essays in Modern Economics*, London: Longmans, 1973.
10. ——— AND VICKREY, W. "Land Use in a Long, Narrow City." *Journal of Economic Theory*, Vol. 3, No. 4 (December 1971), pp. 430-447.

Optimum vs. equilibrium land use pattern and congestion toll

Yitzhak Oron

Department of Economics
Tel-Aviv University

David Pines

Department of Economics and Center for Urban and
Regional Studies
Tel-Aviv University

and

Eytan Sheshinski

Department of Economics
Hebrew University of Jerusalem

This paper presents a comparison between optimum and competitive land use patterns within an urban area. The concept of equilibrium in this paper pertains to five sectors: households, housing producers, composite commodity producers, land transactors and transportation authority. The concept of optimum referred to in this paper is the maximum utility level which can be realized provided that equals are treated equally. This optimum allocation can be supported by a competitive price system if a warranted congestion toll is collected by the transportation authority and redistributed as a lump sum subsidy. If less than the warranted congestion toll is collected, the resulting competitive allocation is distorted and then the competitive city tends to be more suburbanized than the optimum city.

■ This paper presents a model of equilibrium land use pattern in the residential ring of an urban area and compares it to an optimal one. The deviation of the equilibrium land use pattern from the optimal one as a function of the level of congestion toll is examined. In particular, it is examined whether optimal land use pattern is associated with less or more suburbanization in comparison to equilibrium without congestion toll.

In elaborating upon this comparison a general rather than partial equilibrium model is used. It is shown that if the congestion toll does

1. Introduction

Yitzhak Oron received the B.A. from Tel-Aviv University (1968) and the M.Sc.Soc. and the Ph.D. from the Hebrew University of Jerusalem (1970 and 1972, respectively). At Tel-Aviv University, in addition to building urban econometric models, he is conducting research in the differential incidence of taxes and the effect of air pollution on the structure of the urban economic activities.

David Pines received the B.A. (1957), the M.Sc.Soc. (1960), and the Ph.D. (1971) from the Hebrew University of Jerusalem. At Tel-Aviv University he is studying the structure of the housing market, the effect of air pollution on land use pattern, and the application of linear and nonlinear programming to comprehensive regional planning.

Eytan Sheshinski received the B.A. (1960) and the M.A. (1963) from the Hebrew University of Jerusalem and the Ph.D. (1966) from M.I.T. Currently, he is studying general economic theory with a concentration on public economics, at the Hebrew University of Jerusalem.

not internalize all the external effect, then the resulting allocation is inefficient in the sense that relocating households from the periphery to more central locations, keeping their utility constant, can save scarce resources. But this result in itself is not sufficient to assert that the optimal land use pattern is necessarily less suburbanized than the equilibrium one. We turn, therefore, to more specific utility and production functions and solve explicitly for the optimal allocation and the warranted congestion toll. However, it is quite complicated to derive explicitly the misallocation involved by noninternalizing the congestion effect. Therefore, numerical computations are carried out and it is shown, as expected, that an increase in the congestion toll, is associated with a decrease in suburbanization. Though no general conclusion can be derived from the numerical calculations, at least there is an indication that proper pricing of transportation can result in less rather than in more suburbanization.

This result contradicts the conclusions derived recently by Mills regarding the effect of "optimal price of transportation" on the degree of suburbanization.¹ One reason for this contradiction is that Mills builds on a partial equilibrium analysis.

The concept of equilibrium in this paper pertains to five sectors:

- (1) Households
- (2) Housing producers
- (3) Composite commodity producers
- (4) Land transactors
- (5) Transportation Authority (T.A.).

All of the first four agents are competitors: they maximize their objective function [utility in case (1), and profits in cases (2), (3), and (4)], while considering the price system as given. The fifth agent is a regulated nondiscriminating monopoly which determines its price according to some predetermined rule.

In specifying the structure of the economy, it is assumed that the initial holding of a household includes labor hours and a share in the profits of each of the other four sectors. Income is therefore composed of wages and profits. Any change in equilibrium which affects these components affects the budget constraint, not only through the prices of the final goods, but also through the value of the initial holding, i.e., through income. Unfortunately, realization that income is endogenous in the above sense is absent in most of the existing literature on urban land use pattern.

Consider, for instance, the pioneering model of Alonso.² Income of a household is assumed to be exogenously given and is used in determining the bid price function of land (demand price), which in turn is used to determine the equilibrium rent function. In examining the effect of changes in some parameters, income is assumed to remain unaffected. A similar neglect is involved in the recent paper of Solow.³ Though he recognizes that rent is equally redistributed by the government, he treats income as an exogenous variable in comparative statics analysis.

Although Muth, Hochman and Pines, and Mills do not elaborate

¹ In [6].

² In [1].

³ See [11].

on the structure of the economy as suggested above, they succeed in bypassing this difficulty. Muth⁴ assumes that the real-income-constant-price elasticity is constant for a given utility level, and he solves the system up to a constant which depends on real income. Thus, in elaborating upon market equilibrium conditions, he avoids discussion of the income variable. Hochman and Pines⁵ follow Muth. Mills,⁶ too, disregards the income variable by assuming that the demand for housing is exogenously given, i.e., infinitely inelastic with regard to both income and prices.

Skipping over the income variable proves to be useful in deriving very important conclusions regarding the rent and the density gradients. But as will be demonstrated in this paper, income must be explicitly considered in order to examine the difference between equilibrium and optimum land use pattern.

The concept of optimum referred to in this paper is the maximum utility level which can be realized provided that equals are treated equally. In other words, let A be an allocation according to which equal households have equal utility. Let this equal utility be a function defined on the set A . Then "optimum" as referred to here is the allocation which maximizes this function.

The concern with the concept of optimum has been a recent development in urban models. Mills and de Ferranti⁷ determine the allocation of land between housing and transportation that minimizes total transportation costs. Since they assume that the demand for housing is a predetermined magnitude which does not vary with location, their minimization of resources used in transportation implies a maximization of resources available for other uses. Thus, their allocation is efficient. The same holds for the model developed by Solow and Vickrey,⁸ in which the width of the road in a narrow city is determined such that total transportation costs are minimum. The minimization of costs of housing and transportation underlies also the recent approach of Mills to optimum resource allocation in an urban area.⁹

The concept of optimum adopted by Mirrlees¹⁰ is different from the notion discussed so far. His optimum refers to a sum-of-utilities welfare function. As it turns out, his optimum utility levels vary among households. It can be shown, however, that some households are worse off than in our "equal-utility" optimum. In this sense, the unequal treatment of equals implied by the model of Mirrlees cannot be regarded as superior in the Pareto sense to an imposed equal treatment. But when the utility function of the households satisfies the von Neuman-Morgenstern assumptions, and if bundles are distributed by lottery, it can be shown that the allocation suggested by Mirrlees is under certain circumstances optimum according to our notion.¹¹

In the recent paper of Solow,¹² marginal transportation costs are

⁴ In [9, 10].

⁵ In [3].

⁶ In [5].

⁷ In [7].

⁸ In [12].

⁹ In [6].

¹⁰ In [8].

¹¹ See [4].

¹² See [11].

assumed to be proportional to a power of the traffic density. However, it is not elaborated whether this includes the external effect or not. If it does not, then this solution refers to an equilibrium rather than to an optimum. The optimal allocation of land to transportation is, therefore, a second-best solution. If it does include external effect, i.e., including the warranted congestion toll, then income cannot be exogenous, not only because of the redistributed rent but also because of the redistributed toll.

To sum up, the concept of equilibrium is not clearly defined in the existing literature, resulting in ambiguity in the comparative statics analysis. The optimum concept, excluding the one of Mirrlees, is confined to the minimization of social costs.¹³ The concept of Mirrlees may involve unequal treatment of equals, and therefore differs from our notation. We can show that if households are equal in terms of both preferences and initial holdings, then equilibrium without full congestion toll is unambiguously inferior to optimum defined according to our notion, i.e., the optimum utility level of each household according to our notion is higher than in equilibrium, as long as only part of the warranted congestion toll is levied. But equilibrium without the warranted congestion toll is not necessarily unambiguously inferior to optimum according to the definition of Mirrlees, i.e., in his allocation some households may be worse off than in equilibrium.

2. Equilibrium and efficiency of urban spatial structure

■ **Basic assumptions.** The following analysis refers to a monocentric urban form with a radius of S miles. The city is composed of two rings. The first is the Central Business District (CBD), where the production of all the commodities other than housing is carried out. The radius of the CBD is predetermined at one mile. The second is the residential ring, which extends from the boundary of the CBD to the edge of the city. The area of the residential ring is allocated to the transportation network and to housing. The layout of the transportation area is assumed to be predetermined such that the transportation area in a ring at distance x from the center with width dx is $\bar{G}_2(x)dx$. The city is surrounded by an agricultural hinterland where rent is supposed to be uniform, $\bar{R}_A$ units of the composite commodity per unit of area.

Though housing is produced in the residential ring, it is assumed (for convenience of computations) that all the work trips end in the CBD. In other words, workers, like the composite commodity, are transferred from the center to any residential ring without any real costs.

In what follows we first elaborate on the equilibrium conditions of each of the five agents. Then, market equilibrium conditions are formulated to complete the model. Finally, the efficiency of the allocation is checked.

□ **Households.** Households are assumed to be identical in terms of both preferences and their initial holdings. Following Muth and Becker,¹⁴ the utility level of each household is assumed to depend on

¹³ See Mills and de Ferranti [7], Solow [11], and Mills [6].

¹⁴ In [9, 10] and [2], respectively.

housing and on the composite commodity, i.e., it is (directly) independent of location:

$$U(x) = u[H(x), Z(x)],$$

where

$H(x)$ = Quantity of housing consumed by a household located at distance x from the center of the city, and

$Z(x)$ = Quantity of the composite commodity consumed by a household located at distance x from the center.

It is assumed that $u(\cdot)$ is quasi-concave with positive partial derivatives, i.e.,

$$u_H > 0; \quad u_Z > 0,$$

where indices denote partial differentiation with respect to the appropriate variable.

Initial holding of a household is composed of its time and its share in profits of the four other agents. Time is used for leisure, transportation, and production. It is assumed that leisure time is constant. Thus the time spent by each household in work and transportation together is also constant. Households are assumed to have an equal share in profits of land transaction, T.A., and the production activities.

Thus, each household maximizes its utility level subject to a budget constraint:

$$H(x)P_h(x) + Z(x)P_z + \tau(x) = Lw + \psi, \quad (1)$$

where

$P_h(x)$ = Price of housing as a function of distance from the center;

P_z = Price of the composite commodity;

$\tau(x)$ = Costs of transportation at distance x from the center of the city. (These costs include both payment for using the road, if there is any, and the value of time spent in trips.);

$\bar{L}$ = Supply of work hours by a household (for both work and commuting);

w = Wage rate; and

ψ = Household's share in profits from land transaction, profits (losses) of the T.A., and profits from the composite commodity production.

The following are the necessary conditions for equilibrium of the household:

$$u_H/u_Z = P_h/P_z \quad (2)$$

$$HP_h + \dot{\tau} = 0, \quad (3)$$

where a dot signifies a derivative with respect to distance.

Transportation cost includes time spent in commuting plus the toll:

$$\tau(x) = \int_1^x \left[\frac{w}{V(u)} + T(u) \right] du, \quad (4)$$

where

$V(u)$ = Velocity at distance u from the center of the city, and
 $T(u)$ = Toll paid to the T.A. at distance u from the center of the city.

□ **Housing producers.** Housing services are produced by a linear homogeneous function of land and labor, with positive partial derivatives:

$$H^*(x) = h[G_1(x), L_1(x)], \quad (5)$$

where

$H^*(x)$ = Quantity of housing produced at distance x from the center of the city,

$G_1(x)$ = Quantity of land used in the production of housing at distance x from the center of the city, and

$L_1(x)$ = Quantity of labor used in the production of housing at distance x from the center of the city.

Necessary conditions for maximum profits are:

$$h_L P_h = w \quad (6)$$

$$h_G P_h = R(x), \quad (7)$$

where $R(x)$ denotes land rent at distance x from the center of the city.

Net profits derived from housing production are necessarily zero since the production function is homogeneous of the first degree.

□ **Composite commodity producers.** The composite commodity, in contrast to housing and transportation services, is produced only in the CBD. This feature is commonly explained by conditions of increasing returns to scale in urban center production. Actually, this may, in part, account for the existence of such centers.

Production depends on land and labor. Since the area of the CBD is assumed to be fixed, the land input is given. The production function is therefore made to depend explicitly on the variable input alone:

$$Z^* = z(L_2) \quad (8)$$

$$z' > 0, \quad z'' < 0,$$

where

L_2 = Quantity of labor used in the CBD in producing the composite commodity, and

Z^* = Quantity of the composite commodity produced in the CBD.

A necessary condition for maximum is:

$$z' P_z = w. \quad (9)$$

Profits derived in producing the composite commodity equal $z(L_2)P_z - \pi \bar{R}_c - L_2$, where $\bar{R}_c$ is the average value of land in the CBD which is assumed to be exogenous.

□ **Land transactors.** Land is rented from agricultural landlords at a uniform annual cost of $\bar{R}_A$ units of the composite commodity per

unit of land. But the land transactors supplying land at x miles from the center must allocate $G_2(x)/2\pi x$ of the land to the Transportation Authority (privilege of eminent domain). The remaining amount is rented to the highest bidder, provided his demand price exceeds $[2\pi x/\bar{G}_2(x)]\bar{R}_A P_z$. Hence at the very boundary of the city, the following relation must hold:

$$R(S) = \frac{2\pi x}{G_1(x)} \bar{R}_A P_z. \quad (10)$$

The level of rent in the CBD is assumed to be predetermined $\bar{R}_c$. Total profits from land transactions are therefore:

$$\int_1^S [G_1(x)R(x) - 2\pi x \bar{R}_A P_z] dx + \pi \bar{R}_c.$$

This sum is equally distributed among all households.

□ **Transportation authority (T.A.).** The level of service (in terms of velocity) supplied by the T.A. is

$$V(x) = v[E(x), G_2(x)],$$

where

$V(x)$ = Velocity,

$E(x)$ = Number of commuters crossing a ring at distance x from the center, and

$v(\cdot)$ is assumed to satisfy the properties,

$$v_E < 0, \quad v_{EE} < 0, \quad v_G > 0, \quad \text{and} \quad v_{GG} < 0.$$

As mentioned above $\bar{G}_2(x)/2\pi x$ of each unit of urban land located x miles from the center is condemned and allocated to the T.A. for its use. Since $\bar{G}_2(x)$ is a predetermined function, the time spent in crossing a ring at distance x from the center can be represented by:

$$\frac{1}{V(x)} = t[E(x), x].$$

Net income of the T.A. is

$$\int_1^S E(x)T(x)dx.$$

This sum is equally distributed among all households.

□ **Market equilibrium.** The following conditions must hold in equilibrium (the possibility of a corner solution is excluded).

(1) Excess demand for labor is zero:

$$\bar{L}N - \int_1^S (L_1 + tE)dx - L_2 = 0. \quad (12)$$

(2) Excess demand for housing is zero:

$$H^* + MH = 0, \quad (13)$$

where $M(x)$ is the number of households residing beyond x miles

from the center of the city, and

$$M(1) = \bar{N} \quad (14)$$

$$M(S) = 0. \quad (15)$$

(3) Excess demand for the composite commodity is zero:

$$Z^s + \int_1^S MZdx - \pi S^2 \bar{R}_A = 0. \quad (16)$$

The expression $\pi S^2 \bar{R}_A$ is the quantity of the commodity exported to the agricultural hinterland. The agricultural landlords finance these purchases by the rent paid to them by the urban land transactors.

(4) Excess demand for land is zero:

$$2\pi x - \bar{G}_2 - G_1 = 0. \quad (17)$$

(5) Excess demand for transportation is zero:

$$E - M = 0. \quad (18)$$

(6) Total profits accumulated in the economy are distributed among households:

$$\psi = \frac{1}{\bar{N}} \left[Z^s P_z - L_z + \int_1^S (G_1 R + MT) dx - \pi S^2 \bar{R}_A P_z \right]. \quad (19)$$

One of the above equations is redundant. [For instance equation (12) can be derived from equations (1), (4), (11), (16), (18), and (19).] The rest of them can be extended to a standardized system of first-order differential equations in $H(x)$, $Z(x)$, $H^s(x)$, $L_1(x)$, $M(x)$, Z^s , L_z , S , $P_h(x)$, $\tau(x)$, $V(x)$, $R(x)$, P_z , w , and ψ , with the corresponding boundary conditions. One of these, for instance, w , can arbitrarily be chosen.

The characteristics of the resource allocation and the price system can easily be derived, and the results will correspond exactly to what appears already in the existing literature.¹⁵ The differences arise when comparative statics analysis is applied.

□ **Efficiency of resource allocation.** In the following analysis the allocation of land to transportation is assumed to be predetermined. An efficient resource allocation pattern requires that the subjective and objective marginal rate of substituting housing for the composite commodity be equal. This implies:

$$u_H/u_Z = z'/h_L. \quad (20)$$

In addition, it requires that reallocation of a household, keeping its utility level constant, is not associated with saving resources. If one household is relocated from x to $x + \Delta x$, $H(x)/h_L(x)$ labor input is saved in x . On the other hand $H(x)/h_L(x + \Delta x)$ labor is required to produce the same amount of housing in $x + \Delta x$. But in addition, $t[M(x), x]\Delta x$ time is spent in trips by the relocated household and $M(x)t_M\Delta x$ hours by other households. Total social reloca-

¹⁵ See Muth [9], Chapters 2-3.

tion costs are therefore:

$$-\left(\frac{H\dot{h}_L}{h_L^2} + t + Mt_M\right)\Delta x.$$

Efficient allocation implies that total social relocation costs are zero. If they are positive, households should be relocated towards the center, and if negative, households should be relocated towards the periphery.

Is the equilibrium formulated above consistent with efficiency? It follows from (2), (6), and (9) that (20) is satisfied. It remains, therefore, to check whether total social relocation costs are zero. Differentiating (6) with respect to distance it becomes:

$$\dot{h}_L P_h(x) + h_L \dot{P}_h(x) = 0. \quad (21)$$

From equations (3), (4), (11), (18), and (21), it follows that in equilibrium:

$$-H\dot{h}_L/h_L^2 + t + T(x) = 0. \quad (22)$$

Equation (22) implies that total relocation costs are zero *if and only if*, for any x ,

$$T(x) = Mt_M.$$

$T(x)$ is the congestion toll, which must be levied in each location according to the volume of traffic and the capacity of the network. However, under the assumption that demand for commuting is perfectly inelastic, it can be levied according to residential location, x . The warranted congestion toll varies with residential location according to:

$$\int_1^x Mt_M du.$$

It must be emphasized that the warranted congestion toll should correspond to the population distribution M according to optimum allocation rather than to equilibrium without the warranted toll. Hence, in order to determine the warranted congestion toll, the optimal population distribution must be ascertained. This optimum allocation, which corresponds to the above model, is elaborated on in Appendix 1.

■ **Partial equilibrium analysis.** An interesting question is what are the characteristics of the deformation of the structure of the city resulting from a deviation of $T(x)$ from Mt_M . It is especially interesting to examine the effect of an appropriate pricing of transportation on the degree of suburbanization, a problem which has recently been examined by Mills.¹⁶ In what follows, this problem is first solved using partial equilibrium analysis. Such an analysis is shown to imply that the warranted congestion toll causes suburbanization, as Mills asserts, on the basis of numerical computations of a specific model. Second, the limitation of this analysis is discussed, and it is shown that the implications of the partial analysis are not necessarily valid. In the third part, a specific model is solved and numerical computations, which invalidate the implications of the partial analysis, are carried out.

3. The effect of the warranted congestion toll

¹⁶ In [6].

Following the partial equilibrium analysis of Alonso, Muth, and Solow,¹⁷ assume that income and the price of the composite commodity are both predetermined. Assume that only a fraction of the congestion toll is levied. Then, imposing the warranted congestion toll will improve the allocation of resources, resulting thus in higher utility level of every household (households are identical and their utility in equilibrium must therefore be the same). Hence the utility of households located in the very boundary of the CBD also increases.

Households in the very boundary of the CBD do not spend income on transportation cost. Their money income and the price of the composite commodity are assumed to be predetermined. Hence, the only way that the budget constraint can be modified to allow for the increase in welfare is through decline in the price of housing. This must lead to an increase in the per household consumption of housing.

The production function of housing is assumed to be linear homogeneous and the wage rate is predetermined. Under these assumptions the equality of the value of marginal product to factor cost (land and labor) implies that the average product of land in the production of housing is an increasing function of the price of housing. Thus, since the price of housing declines near the boundary of the CBD, the average product of land in the production of housing declines there. Hence, per household consumption of housing increases and housing-land ratio declines near the boundary of the CBD, resulting in a decrease in net density (household per residential area). Moreover, since the residential area is predetermined [$\bar{G}_2(x)$ is assumed to be exogenous and therefore $x - G_2(x)$], gross density (household per unit of total area) also declines.

Of course, the price of housing cannot decrease everywhere, since if it does the density decreases everywhere and the city size must also shrink. In this case the supply of housing falls short of demand. The conclusion is that somewhere the price must increase. Assuming continuity of all the functions involved, gross density must decline in some range around the CBD and increase somewhere in more distant locations.

Note that in contrast to Mills, the above conclusion is derived without referring to specific utility, congestion, and production functions. Moreover, no numerical calculations are necessary. But the shortcoming of this analysis is that income and the price of the composite commodity are assumed to be predetermined. If the imposition of the warranted congestion toll affects these variables, the above analysis is not necessarily valid anymore.¹⁸ As a matter of fact it

¹⁷ In [1], [9, 10], and [11], respectively.

¹⁸ As a matter of fact Mills' analysis is designed to compare optimum allocation with competitive allocation when no congestion toll is charged and when production of transportation, G_2 , is predetermined. Methodologically this analysis can be subdivided into two distinct stages. First, the impact of congestion toll is examined (so far this is the subject of the discussion in the text). Second, the impact of efficient allocation of resources to transportation is checked. The results of the second stage analysis clearly depend on the initial pattern of the transportation production. Consider for instance two extreme cases. According to the first, all the land is devoted to transportation around the CBD, and this "transportation ring" extends far enough (e.g., beyond the boundary of the city in optimal allocation). According to the second, no land is allocated to transportation in the fringe of the city (which is within the boundary of the city in

follows from equation (22) that if $T(x) < Mt_M$, then it is desirable to relocate people towards the center. Does this imply that in the above equilibrium the city is too dispersed relative to optimal resource allocation as intuition tends to indicate? There is no definite answer. At the optimum, real income should be higher. Therefore, housing consumption is higher, which may result in a bigger city. This is more likely the higher is the income elasticity of the demand for housing. It is difficult, however, to state general conclusions. We turn therefore to specific model and numerical computations.

□ **A general equilibrium approach—a specific model.** In this section the effect of the warranted congestion toll is examined by using a general equilibrium model. But the model solved is specific in the sense that certain utility and congestion functions are assumed. The optimum solution is derived analytically. The solution for equilibrium is derived by numerical computations.

The effect of the warranted congestion toll depends of course on the initial layout of the transportation network. For convenience of the computations, it is assumed here that the initial layout of the transportation network is according to the optimal resource allocation.¹⁹

In solving the system, the radius of the city has been assumed to be predetermined, i.e., the boundary condition (10) is disregarded. But this is only a working assumption. We evaluate the effect on suburbanization by the effect on the level of rent at the boundary of the city. If the equilibrium rent increases at the boundary with a change in the exogenous variable, then the change is associated with suburbanization. This result can be expected if the predetermined radius of the city was the equilibrium one, i.e., one which satisfies equation (10) above.

Suppose that both housing and transportation are produced only by land:

$$H^s(x) = G_1(x) \quad (23)$$

$$T^s(x) = G_2(x), \quad (24)$$

where

$T^s(x)$ = Quantity of transportation services produced at distance x .

Equality of supply and demand for land implies:

$$G_2(x) = 2\pi x - G_1(x). \quad (25)$$

Suppose that the velocity function is:

$$\frac{1}{V(x)} = \frac{M(x)}{\bar{\alpha} G_2(x)}, \quad (26)$$

optimum allocation). It is clear that in the first case optimum implies a more compact city than equilibrium. The opposite is true in the second case. These extreme cases are sufficient for illustrating that the impact of efficient resource allocation to transportation depends on the initial pattern of transportation production. The problem discussed in the text is therefore whether the impact of congestion toll can, *a priori*, be ascertained.

¹⁹ See footnote 18 above.

where

$\bar{\alpha}$ = A technological coefficient, and

$V(x)$ = Velocity at distance x from the center of the city.

The composite commodity is produced in the CBD with labor services only:

$$z = \bar{\alpha}L_c, \quad (27)$$

where $\bar{\alpha} > 0$ is a technological coefficient.

The utility function is assumed to be a product of the quantities of housing services and the composite good²⁰:

$$U(x) = H(x)Z(x). \quad (28)$$

Suppose that it is required to determine the resource allocation which will yield a maximum level of equal utility. In other words, of all the allocations which imply equal utility to everyone, we look for the one which maximizes this utility. Control theory is used to solve the problem. The sum of utilities is maximized subject to labor and land constraints and an additional constraint which requires that utilities are equal. Skipping some technical steps, the necessary conditions for such a maximum are the following:

$$G_1(x) = -\frac{\pi}{2A_1} + \pi x + \frac{1}{4A_1^2} e^{A_4+2A_1x} \quad (29)$$

$$G_2(x) = \frac{\pi}{2A_1} + \pi x - \frac{1}{4A_1^2} e^{A_4+2A_1x} \quad (30)$$

$$M(x) = \frac{\bar{\alpha}}{2} A_1 e^{-A_2-A_1x} G_2(x) \quad (31)$$

$$-\dot{M}(x) = \frac{\bar{\alpha}}{2} A_1^2 e^{-A_2-A_1x} G_1(x) \quad (32)$$

$$Z(x) = \frac{\bar{\alpha}}{2\bar{\alpha}} e^{-A_2-A_1x}, \quad (33)$$

where A_1 , A_2 , and A_4 are constants to be determined by the following equations:

$$e^{A_4} = \pi \left(\frac{1}{2A_2} + \bar{S} \right) 4A_1^2 e^{-2A_1\bar{S}} \quad (34)$$

$$e^{A_2} = \frac{\bar{\alpha}\pi}{2} A_1 e^{-A_1\bar{\epsilon}} \frac{1}{\bar{N}} \left[\frac{1}{2A_1} + \bar{\epsilon} - e^{2A_1(\bar{\epsilon}-\bar{S})} \left(\frac{1}{2A_1} + \bar{S} \right) \right] \quad (35)$$

$$\begin{aligned} \frac{\bar{L}}{\bar{N}} \bar{\alpha} = & \frac{e^{2A_1\bar{\epsilon}}}{\pi A_1} \left[\frac{1}{2A_1} + \bar{\epsilon} - e^{2A_1(\bar{\epsilon}-\bar{S})} \left(\frac{1}{2A_1} + \bar{S} \right) \right. \\ & \left. \times \left(e^{-2A_1\bar{\epsilon}} - e^{-2A_1\bar{S}} + \frac{1}{2A_1} (e^{-2A_1\bar{\epsilon}} - e^{-2A_1\bar{S}}) \right) \right], \quad (36) \end{aligned}$$

where $\bar{\epsilon}$ is the radius of the CBD.

²⁰ Or any other monotonic transformation of this function.

The maximized utility level is:

$$u_0 = -\frac{\bar{a}}{\bar{a}} A_1^{-2}. \quad (37)$$

It was shown in Section 2 that a full congestion toll is required if the equilibrium allocation is to be also an optimal allocation, i.e.,

$$F(x) = \int_{\bar{x}}^x T(u) du, \quad (38)$$

where $T(u) = Mt_M$.

According to the specific model discussed in this section,

$$t[M(x), x] = \frac{M(x)}{\bar{a}G_2(x)} = \frac{1}{V(x)},$$

and, therefore,

$$t_M M(x) = \frac{M(x)}{\bar{a}G_2(x)}. \quad (39)$$

Thus, equations (31) and (39) yield:

$$t_M M(x) = \frac{A_1}{2} e^{-A_2 - A_1 x}. \quad (40)$$

And equations (38) and (40) imply:

$$F(x) = \int_{\bar{x}}^x \frac{A_1}{2} e^{-A_2 - A_1 u} du = \frac{A_1}{2} e^{-A_2} \left[-\frac{1}{A_1} e^{-A_1 u} \right]_{\bar{x}}^x.$$

Hence, the optimal transportation charge is

$$F(x) = \frac{e^{-A_2}}{2} (e^{-A_1 \bar{x}} - e^{-A_1 x}), \quad (41)$$

and we find

$$T(x) = \frac{1}{V(x)} = t_M M(x) = \frac{A_1}{2} e^{-A_2 - A_1 x}. \quad (42)$$

It is interesting to examine how the warranted congestion toll should vary with some specific changes in parameters. Suppose for instance that both the size of the population and the transportation efficiency $\bar{a}$ are increased by the same coefficient λ :

$$\bar{N}_1 = \lambda \bar{N}$$

$$\bar{a}_1 = \lambda \bar{a}.$$

According to (34) through (37), A_1 , A_2 , and A_4 remain the same. Moreover, according to (29) through (33), velocity, allocation of land, and quantity of the composite goods consumed by households at distance x , also remain the same. However, both the number of commuters crossing any given ring at distance x and the number of households residing in x increase at a rate of λ . Therefore, the quantity of housing consumed by each household decreases by the factor $1/\lambda$.

The above two simultaneous changes do not affect the warranted congestion toll which still remains the same. This can easily be verified

by checking the effect on (42): nowhere does the velocity change.

So far it was demonstrated that if the warranted congestion toll is levied, the resulting allocation is optimal relative to an equal distribution. But, with no restrictions on the form of the functions, we were unable to obtain conclusions (Section 2) regarding the effect of the level of the toll on the urban form.

The purpose of this section is to verify, by using the specific assumptions regarding the utility and the production functions, the effect of the level of the toll on the urban form.

Suppose that γ percent of the warranted congestion toll is charged in each location. The resulting allocation can be derived from the solution of the following three differential equations:

$$-\dot{M}(x) = \frac{\bar{L} + \psi}{2A_1^2} G_1(x) - \frac{\gamma + 1}{2\bar{\alpha}A_1^2} G_1(x) \int_{\bar{x}}^x \frac{M(u)}{G_2(u)} du \quad (43)$$

$$\psi = -\frac{1}{\bar{N}} \int_{\bar{x}}^s \frac{A_1^2}{G_1(x)} \dot{M}^2(x) + \frac{\gamma}{\bar{\alpha}\bar{N}} \int_{\bar{x}}^s \frac{M^2(x)}{G_2(x)} dx \quad (44)$$

$$A_1 = (u_1/\bar{\alpha})^{\frac{1}{2}}, \quad (45)$$

where u_1 is the resulting equilibrium utility level.

Equation (43) is a second-order differential equation which we fail to solve analytically. Solving it numerically for arbitrary parameters, we found that when the warranted congestion toll is imposed, the density in the center increases and the urban rent in the very boundary of the city decreases. This indicates that the city becomes less suburbanized. This trend exists also when only a portion of the warranted congestion toll is effective, i.e., γ percent of the warranted congestion toll. The higher the γ , the less urbanized the city is. (See Appendix 2.)

Appendix 1

Optimum allocation when equals are equally treated

■ In the following discussion the optimum allocation referred to in this paper will be elaborated on, and the warranted congestion toll will be derived. It will be shown that if there exists such an optimum, and the warranted congestion toll is levied and then distributed as a lump sum subsidy, there must exist a price system which can support a competitive allocation which corresponds exactly to the optimum allocation.

Using the notation of the text, the optimum allocation referred to in this paper is the solution of the following problem:

Maximize W

Subject to:

$$W - u[H(x), Z(x)] = 0$$

$$\dot{M}H(x) + h[G_1(x), L_1(x)] = 0$$

$$G_1(x) + \bar{G}_2(x) - 2\pi x = 0$$

$$\int_1^s \{L_1(x) + M(x)t[M(x), x]\}dx + L_2 - \bar{L}\bar{N} = 0 \quad (A1)$$

$$\int_1^S -\dot{M}(x)Z(x)dx + \pi S^2 R_A - z(L_2) = 0, \quad (A2)$$

and the boundary conditions $M(1) = N$ and $M(S) = 0$.

The utility function can be rewritten as

$$U(x) = u[-H^S/\dot{M}(x), -C/\dot{M}(x)],$$

where $C = -Z\dot{M}$.

Since it was assumed that the marginal utility of each commodity, i.e., U_H and U_Z , increases with its quantity, one can define a function $f(\cdot)$ such that

$$\dot{M} = -f(H^S, C, U). \quad (A3)$$

$f(\cdot)$ can be shown to be homogeneous of the first degree with respect to H^S and C .

By using equation (A3), the Hamiltonian can be defined as:

$$\mathcal{L} = [W - \pi(x)]f(H^S, C, W) - \lambda_2 C - \lambda_1(L_1 + tM),$$

where $\pi(x)$ is the auxiliary function corresponding to $M(x)$, and λ_1 and λ_2 are the Lagrangian multipliers of constraints (A1) and (A2), respectively.

Necessary conditions for solution of the above problem are:

$$\frac{\partial \mathcal{L}}{\partial L_1} = (W - \pi)f_H h_L - \lambda_1 = 0 \quad (A4)$$

$$\frac{\partial \mathcal{L}}{\partial C} = (W - \pi)f_C - \lambda_2 = 0 \quad (A5)$$

$$\frac{\partial \mathcal{L}}{\partial M} = \lambda_1(t + Mt_M) = \dot{\pi} \quad (A6)$$

$$\lambda_2 z' - \lambda_1 = 0. \quad (A7)$$

Equations (A1) through (A7) are to be used to solve for W , L_2 , $\pi(x)$, $C(x)$, $M(x)$, $L_1(x)$, and λ_2 in terms of λ_1 .

Suppose now that W^* , L_2^* , $\pi^*(x)$, $C^*(x)$, $M^*(x)$, $L_1^*(x)$, λ_2^* , and λ_1^* solve the above problem. Then it can be shown that the resource allocation $[L_2^*, C^*(x), L_1^*(x), M^*(x)]$ can be supported by the following price system:

$$w = \lambda_1^* \quad (A8)$$

$$P_z = \lambda_2^* \quad (A9)$$

$$P_h = (W^* - \pi)f_H \quad (A10)$$

$$R = P_h h_G^* \quad (A11)$$

$$T = M^* t_M^*. \quad (A12)$$

To demonstrate this assertion it is only necessary to show that equations (1), (2), (3), (6), (7), (9), and (10) hold as long as the initial holding of any household is $\bar{L}$ units of labor plus an entitlement to equal share in profits accumulated in the economy [i.e., equation (19) is satisfied].

TABLE A1

SET I

γ	RESIDENTIAL DENSITY AT THE BOUNDARY OF THE CBD*	URBAN RENT AT THE BOUNDARY OF THE CITY	TOTAL TIME SPENT IN COMMUTING	U_o
0.0	0.74449	13247	120308	54.446
0.1	0.74923	13121	120098	54.453
0.2	0.75396	13060	119889	54.459
0.3	0.75868	12968	119681	54.465
0.4	0.76345	12907	119465	54.467
0.5	0.76815	12787	119260	54.472
0.6	0.77290	12698	119048	54.473
0.7	0.77756	12610	118848	54.477
0.8	0.78226	12553	118642	54.479
0.9	0.78696	12467	118435	54.480
1.0	0.79136	12358	118277	54.500
*IN SQUARE FEET PER CAPITA.				

TABLE A2

SET II

γ	RESIDENTIAL DENSITY AT THE BOUNDARY OF THE CBD*	URBAN RENT AT THE BOUNDARY OF THE CITY	TOTAL TIME SPENT IN COMMUTING	U_o
0.0	0.13260	10860	75391	71.859
0.1	0.13327	10790	75280	71.861
0.2	0.13391	10726	75175	71.868
0.3	0.13456	10664	75070	71.874
0.4	0.13522	10617	74961	71.875
0.5	0.13586	10555	74859	71.881
0.6	0.13652	10509	74752	71.882
0.7	0.13717	10433	74646	71.884
0.8	0.13781	10382	74542	71.887
0.9	0.13846	10328	74438	71.888
1.0	0.13910	10285	74336	71.890
*IN SQUARE FEET PER CAPITA.				

TABLE A3

SET III

γ	RESIDENTIAL DENSITY AT THE BOUNDARY OF THE CBD*	URBAN RENT AT THE BOUNDARY OF THE CITY	TOTAL TIME SPENT IN COMMUTING	U_o
0.0	0.16907	13428	106583	55.994
0.1	0.17003	13333	106407	55.997
0.2	0.17098	13239	106241	56.002
0.3	0.17192	13178	106074	56.007
0.4	0.17287	13086	105903	56.008
0.5	0.17381	13026	105739	56.013
0.6	0.17476	12936	105570	56.014
0.7	0.17570	12847	105403	56.015
0.8	0.17664	12789	105239	56.017
0.9	0.17757	12702	105077	56.018
1.0	0.17843	12678	104955	56.040
*IN SQUARE FEET PER CAPITA				

In order to derive equation (1), we first use the property that $f(\cdot)$ is linear homogeneous with respect to H and Z to obtain:

$$-\dot{M} = f_C C + f_H H^S. \quad (A13)$$

Then we use equations (4), (11), (19), (A1), (A2), (A5), (A6), (A8), (A9), (A10), (A12), and (A13) to derive equation (1).

It follows from the properties of $f(\cdot)$ that $u_H/u_Z = f_H/f_C$. This and equations (A5), (A9), and (A10) are used to derive equation (2).

Equation (3) follows from equations (1), (2), and the property that the utility does not vary with distance.

Equation (6) follows from equations (A4) and (A10). Equation (7) is given by equation (A11); equation (9) follows from equations (A7) and (A9), and equation (10) follows from the transversality condition $\mathcal{L}(S) = 2\lambda_2 \pi x$.

Hence, if there exists an optimum allocation there must also exist a price system which includes a congestion toll $M^* t_M^*$.

Appendix 2

■ There are three sets of parameters which were selected.

□ Set I. (See Table A1.)

$$\begin{array}{lll} \bar{\epsilon} \text{ (radius of the CBD)} = 2 & \bar{S} = 15 & \bar{a} = 1 \\ \bar{N} = 100,000 & \bar{\alpha} = 80,000 & \end{array}$$

For this set of parameters the optimal allocation of land is:

$$G_1(x) = \frac{\pi}{0.030275} + \pi x + 95.877e^{0.030275x}.$$

□ Set II. (See Table A2.)

$$\begin{array}{lll} \bar{\epsilon} = 2 & \bar{S} = 15 & \bar{a} = 1 \\ \bar{N} = 80,000 & \bar{\alpha} = 100,000 & \end{array}$$

$$G_1(x) = -\frac{\pi}{0.02858} + \pi x + 172e^{0.02858x}.$$

□ Set III. (See Table A3.)

$$\begin{array}{lll} \bar{\epsilon} = 2 & \bar{S} = 15 & \bar{a} = 1 \\ \bar{N} = 100,000 & \bar{\alpha} = 100,000 & \end{array}$$

$$G_1(x) = -\frac{\pi}{0.02669} + \pi x + 110.49e^{0.02669x}.$$

References

1. ALONSO, W. *Location and Land Use*. Cambridge: Harvard Univ. Press, 1965, Chapters 2, 4, and 5.
2. BECKER, G. S. "The Theory of Allocation of Time." *Economic Journal*, Vol. 75, No. 298 (September 1965), pp. 493-517.
3. HOCHMAN, O. AND PINES, D. "Competitive Equilibrium of Transportation and Housing in the Residential Ring of an Urban Area." *Environment and Planning*, Vol. 3 (1971), pp. 51-62.

4. LEVHARI, D., ORON, Y., AND PINES, D. "In Favor of Lottery in Cases of Non-convexity." Working Paper No. 10, Center for Urban and Regional Studies, Tel-Aviv University, May 1972.
5. MILLS, E. S. "An Aggregate Model of Resource Allocation in a Metropolitan Area." *The American Economic Review*, Vol. 57, No. 2 (May 1967), pp. 197-210.
6. ———. *Studies in the Structure of the Urban Economy*. Baltimore: Johns Hopkins Press, 1972, Chapters 5-8.
7. ——— AND DE FERRANTI, D. M. "Market Choices and Optimum City Size." *The American Economic Review*, Vol. 61, No. 2 (May 1971), pp. 340-345.
8. MIRRELES, J. A. "The Optimum Town." *Swedish Journal of Economics*, Vol. 74, No. 1 (March 1972), pp. 114-135.
9. MUTH, R. F. *Cities and Housing*. Chicago: Univ. of Chicago Press, 1969, Chapters 2-4.
10. ———. "The Spatial Structure of the Housing Market." *Papers and Proceedings of the Regional Science Association*, Vol. 7 (1961).
11. SOLOW, R. M. "Congestion, Density and the Use of Land in Transportation." *Swedish Journal of Economics*, Vol. 74, No. 1 (March 1972), pp. 161-173.
12. ——— AND VICKREY, W. S. "Land Use in a Long Narrow City." *Journal of Economic Theory*, Vol. 3, No. 4 (December 1971), pp. 430-447.

The optimum factory town

Avinash Dixit

Balliol College
Oxford University

This paper studies the determination of the optimum size and arrangement of a monocentric city. The major consideration is the trade-off between economies of scale in production in the central business district and the diseconomies of congestion in commuter transport. Planned optimum allocations and their decentralized implementation by suitable taxes and transfers are discussed. Rent and density profiles and land use are solved in closed form, and complete numerical solutions are obtained for particular parameter values. It is shown that negative exponential rent and density profiles may not be very good approximations, and that cities of more than a million inhabitants are difficult to justify in the framework considered.

■ The general conceptual framework of the theory of monocentric cities is too well-known to need another airing.¹ In this paper I offer improvement on earlier work at several specific points. I shall indicate these by comparing this paper with some recent ones.

The most serious defect common to several of these papers (e.g. those of Solow and of Oron, Pines, and Sheshinski²) is their assumption of nonincreasing returns to scale in production in the central business district. This destroys all rationale for organizing production in a city, as dispersed production will save transport costs without losing output. To preserve the urban structure, we must then impose the condition that all output has to be shipped to a point market place, which is an arbitrary and unsatisfactory way out. The role of increasing returns in understanding urban problems is clearly stated by Koopmans,³ and general theorems relating city size to increasing returns are proved by Mirrlees and Starrett.⁴ I shall be concerned with more specific cases which can yield approximate numerical results, and shall therefore use particular functional forms and study the role of increasing returns by varying suitable parameters in a production function.

Avinash Dixit is Lord Thomson of Fleet Fellow at Balliol College, Oxford. He received the B.S. from the University of Bombay (1963), the B.A. and the M.A. from Cambridge University (1965 and 1969, respectively), and the Ph.D. from M.I.T. (1968). His current research centers on theoretical urban economics and development economics.

This paper was written while the author was visiting the Massachusetts Institute of Technology. He would like to thank the Department of Economics for their hospitality, the National Science Foundation for financial support, Robert Solow and Nicholas Stern for several valuable theoretical suggestions, Jerome Rothenberg and William Wheaton for practical ones, and Christopher Cavanagh for programming some of the computations.

¹ Good textbook expositions are in Mills [3] and Muth [6].

² In [12] and [8], respectively.

³ In [1], p. 154.

⁴ In [5] and [14], respectively.

1. Introduction

Models which have allowed for increasing returns have treated transport in a cursory way. My main theme will be that optimum city size is determined by the balance between economies of scale in production and diseconomies in transport, congestion being an important element in the latter. It is obvious that this leads directly to the analysis of land use in the residential belt. The voluminous literature on this subject is unsuited for the aim I have in mind. Mills and de Ferranti and Solow,⁵ for example, avoid the problem by making special assumptions about land allocation for residence and transport. Riley discusses the general principles involved, and Oron et al. consider an unduly restrictive form of the transport cost function.⁶ Again, my aim will be to parametrize the problem, but to retain enough richness to allow at least a first approximation to reality.

Existing parametric models fail to be realistic in many ways. Solow and Oron et al.⁷ assume that costs per unit of transport are directly proportional to the traffic density, and that consumers divide their disposable income equally between housing and other expenditures. Both of these are far from reality, the second particularly so. Further, both turn out to affect the solution in the same direction. Even worse, the criterion of optimality chosen by Oron et al. is such as to reinforce this effect. It is therefore not possible to expect the special assumptions to neutralize one another, and better models are essential. Numerical work of Solow⁸ removes many of these restrictions from the problem of equilibrium allocation. I shall concentrate on the corresponding optimum problem. Using a fairly rich parametric specification, I shall obtain the entire optimum residential belt allocation in closed form, and then use the end conditions and production relations to solve the complete model numerically. This has the double advantage of presenting as many results as possible more vividly in explicit functional form, and of enabling us to do more and simpler numerical calculation at the end.

Finally, I shall consider the problem of optimum allocation in the fullest sense. Most earlier work has been partial in holding some relevant variables fixed at arbitrary values. Mills and de Ferranti fix the size of the business district and do not introduce production at all. Oron et al. fix the central business district, the city boundary, and the population. Mirrlees fixes the land use requirements for transport. I shall allow all these to be policy variables.

2. The model

■ The general structure of the model can be described as follows. Suppose a very large population P is to be split into self-sufficient towns of size N each, where P and N refer to households. Suppose this arrangement yields a value $W(P, N)$ of some welfare criterion. We wish to choose N to maximize this. With this aim in mind, I shall first study the optimum organization of a town with a given number N of households to evaluate the function W , then find N to maximize it, and finally study the dependence of this whole arrangement on the parameters of the model.

⁵ In [4] and [11], respectively.

⁶ In [10] and [8], respectively.

⁷ In [11] and [8], respectively.

⁸ In [13].

Much of the model will be built on the lines of Oron et al. Let the town of radius s have a concentric business district of radius c and area $E = \pi c^2$. Gross output produced by L man-hours working on this area is

$$Z = AL^\alpha E^\beta, \quad (1)$$

where α and β each lie in $(0, 1)$ and their sum exceeds one. Thus there are increasing returns to scale and diminishing returns to each factor. Of this output, an amount $\pi \rho s^2$ goes to pay rent to agricultural landlords who once owned the land. Also, amount $Q = \gamma E$ is needed as an input for production. This can be thought of as implicit rental costs of plant and equipment, or as public services. When the government controls the use of land as in this model, either interpretation will do. The implicit assumption that this input is perfectly complementary to land is very special, but it is better than nothing. With this, an amount $(Z - \pi \gamma c^2 - \pi \rho s^2)$ is available for urban consumption.

The residential ring between c and s contains N households. Define $N(x)$ as the number between x and s , so $-N'(x)dx$ is between x and $x + dx$. If each household at x gets $z(x)$ of the consumer good, we have

$$\int_c^s z(x)[-N'(x)]dx = Z - \pi \gamma c^2 - \pi \rho s^2. \quad (2)$$

In the thin ring between x and $x + dx$, $2\pi G_1(x)dx$ units of land are used for housing and $2\pi G_2(x)dx$ for traffic, with

$$G_1(x) + G_2(x) = x. \quad (3)$$

Now each household at x gets $h(x) = -G_1(x)/N'(x)$ units of housing, and from this and $z(x)$ of the consumer good, it derives utility

$$u(x) = z(x)^{\sigma/(1+\sigma)} h(x)^{1/(1+\sigma)}. \quad (4)$$

This means that if the household chooses its consumption given income and prices, it would divide its income between consumer goods and housing in proportions of $\sigma:1$. Solow and Oron et al. assume $\sigma = 1$. To be more realistic, σ must be at least three; if we take a minimal definition of housing so that many aspects of housing are aggregated with the consumer good, then σ can be as high as nine.

All transport in the town is commuter traffic, and a worker who lives at x has to spend $T(x)$ hours commuting, where

$$T'(x) = \tau + \delta[N(x)/G_2(x)]^k. \quad (5)$$

This says that costs per unit distance at x consist of a pure distance cost element and a congestion effect. Later I shall specialize to the case of pure congestion costs, i.e., $\tau = 0$. Again, $k = 1$ is a common assumption, but a higher value will be more realistic; Vickrey⁹ believes k will exceed two.

I shall assume that each household can supply a fixed number H of hours for work and commuting taken together, and that there are no transport costs within the business district. Then the labor hours available are given by

$$L = NH - \int_c^s T(x)[-N'(x)]dx.$$

⁹ In [16].

On integrating by parts, using $T(c) = 0 = N(s)$, and (5),

$$L = NH - \int_c^s [\tau N(x) + \delta N(x)^{1+k}/G_2(x)^k] dx. \quad (6)$$

Of the variables introduced so far, c , s , and the functions $G_1(x)$, $G_2(x)$, $z(x)$, and $N(x)$ will be policy variables. This completes the description of the model.

It remains to decide on a criterion of optimality. Mirrlees maximizes a general additively separable welfare function, and finds that all residents will get equal utility only in very special cases.¹⁰ The extent of inequality is not examined in any detail. Now unequal treatment of identical people is ethically unappealing. Besides, it is difficult to conceive a viable social system which begins with identical people and then introduces any significant degree of inequality among them. For these reasons, we should like to have some idea of the extent of inequality involved. Oron et al. carry a concern for equality to the extreme by imposing a constraint that all residents have equal utility. This is quite strong, and we should like to see how it relates to the approach of Mirrlees.

I shall use an objective function which enables us to make such comparisons in terms of a parameter. This is a familiar constant elasticity function, and the extent of inequality depends on the elasticity. Remember that $u(x)$ in (4) was homogeneous of degree one. My maximand will be

$$J = \int_c^s -u(x)^{-m} [-N'(x)] dx, \quad (7)$$

where $m > 0$. A higher m will lead to less inequality in the optimum allocation. In the limit as $m \rightarrow \infty$, the objective becomes equivalent to maximizing $\inf u(x)$, the criterion made famous by Rawls.¹¹ Here I shall present the case of a general finite m , and draw conclusions for the Rawlsian case by taking limits. Of course this can lead to pitfalls, and I have checked the validity of all the remarks I shall make by a separate direct solution of the Rawlsian case. It turns out that utility is equally distributed in that case, and therefore it is equivalent to the criterion of Oron et al.

My emphasis throughout will be on the centrally planned optimum. It is, however, interesting to note that it can be decentralized by using suitable taxes and transfers, and some of the results will be easier to interpret and compare with earlier work in terms of such decentralized equilibrium. The problem has been considered in detail for similar models by Mirrlees and by Oron et al., and their proofs are easily adapted for my model. I shall therefore only state the rules for decentralization. Labor in the central business district is paid its marginal product, and land is the residual claimant. This works so long as we do not have increasing returns to labor alone. Commuters must be charged congestion tolls to equate the private and social costs of transport. Total costs per unit distance of transport at x are $N(x)T'(x)$. Given the road structure $G_2(x)$, using (5) and differentiating with respect to $N(x)$, the marginal costs are

¹⁰ In [5].

¹¹ In [9].

$\tau + \delta(k+1)[N(x)/G_2(x)]^k$. The proper congestion toll per unit distance at x is then the excess of this over the private cost namely $\delta k[N(x)/G_2(x)]^k$, which equals $k[T'(x) - \tau]$. Integrating, we see that a commuter at x should pay an amount $k[T(x) - \tau(x - c)]$ in wage units for his two-way commuting ticket. Of course he pays a private cost $T(x)$ in lost wages as well. Finally, lump sum subsidies as a function of distance can be so arranged as to lead to the optimum location and housing decisions. In the Rawlsian case where all households have equal utility, the market realization will provide them the same opportunity set, so the lump sum subsidy will be independent of the distance. The most useful information from the market realization is the rent function. Once we know the full optimum, this can easily be calculated as the ratio of marginal utilities, and written as $r(x) = u_h(x)/u_z(x) = z(x)/[\sigma h(x)]$.

■ Let $v(x) = u(x)^{-m}$, and write (4) as

$$v(x)^{-(1+\sigma)/m} N'(x) + z(x)^\sigma G_1(x) = 0. \quad (4')$$

Now we can maximize (7) subject to (1)–(3), (4'), and (6). Using a Lagrangian multiplier for each constraint, we get

$$\mathcal{L} = \int_0^s \mathcal{L} dx + \eta[Z - \pi\gamma c^2 - \pi\rho s^2] + \theta(NH - L) + \omega[AL^\alpha E^\beta - Z], \quad (8)$$

where $\mathcal{L}$ is the Lagrangian,

$$\begin{aligned} \mathcal{L} = & v(x)N'(x) + \lambda(x)[v(x)^{-(1+\sigma)/m} N'(x) + z(x)^\sigma G_1(x)] \\ & + \mu(x)[x - G_1(x) - G_2(x)] + \eta z(x)N'(x) \\ & - \theta[\tau N(x) + \delta N(x)^{1+k}/G_2(x)^k]. \end{aligned} \quad (9)$$

The policy variables are $c, s, G_1(x), G_2(x), z(x)$, and $N(x)$, and it is convenient to regard $v(x), Z$, and L as further choice variables subject to the appropriate constraints. Total population is fixed for now; its choice will be considered later.

First-order conditions for the choice of the five functions are:¹²

$$N'(x) - \frac{1+\sigma}{m} \lambda(x) v(x)^{-(1+\sigma)/m-1} N'(x) = 0,$$

$$\lambda(x) z(x)^\sigma - \mu(x) = 0,$$

$$-\mu(x) + k\delta\theta[N(x)/G_2(x)]^{1+k} = 0,$$

$$\sigma\lambda(x)z(x)^{\sigma-1}G_1(x) + \eta N'(x) = 0,$$

and

$$\begin{aligned} \frac{d}{dx} [v(x) + \lambda(x)v(x)^{-(1+\sigma)/m} + \eta z(x)] \\ = -\theta\tau - (1+k)\delta\theta \left[\frac{N(x)}{G_2(x)} \right]^k. \end{aligned}$$

For Z and L , we have

$$\eta - \omega = 0$$

$$-\theta + \omega\alpha Z/L = 0.$$

3. Optimum allocation

¹² Mathematical details of the optimization are available upon request from the author.

Finally, the choice of c and s is determined by the transversality conditions,

$$-\mathcal{H}(c) + \theta HN'(c) + 2\pi\eta c(\beta Z/E - \gamma) = 0, \quad \backslash$$

and

$$\mathcal{H}(s) - 2\pi\eta\phi s = 0,$$

where $\mathcal{H} = \mathcal{L} - N'\partial\mathcal{L}/\partial N'$ is the Hamiltonian.

I shall use the last four conditions only in numerical calculations. The first five give several interesting qualitative results. We get

$$v(x) = \text{const} \cdot z(x), \quad (10)$$

$$u(x) = \text{const} \cdot z(x)^{-1/m}, \quad (10')$$

$$N(x)/G_2(x) = \text{const} \cdot z(x)^{a/k}, \quad (11)$$

$$-N'(x)/G_1(x) = \text{const} \cdot z(x)^{\sigma+(1+\sigma)/m}, \quad (12)$$

and

$$z'(x) = -\text{const} [\tau + \delta \cdot \text{const} \cdot z(x)^a], \quad (13)$$

where the constants, all different, are positive, and depend on the parameters and on θ , on η , but not, of course, on x . Also,

$$a = \frac{k(1+m)(1+\sigma)}{(1+k)m}. \quad (14)$$

Finally, using (3), we get

$$\frac{d}{dx} [z(x)^a N(x)] + \text{const} \cdot x z(x)^{a+\sigma+(1+\sigma)/m} = 0. \quad (15)$$

Equation (13) is easy to solve explicitly in special cases. First, if congestion costs are negligible, $\delta = 0$ and $z(x)$ falls linearly. This is a less interesting case, and I shall not consider it any further. Next, if pure distance costs are negligible: $\tau = 0$. This is the case considered for the rest of the paper. Finally, if $a = 1$, we get

$$\tau + \delta c_1 z(x) = c_2 e^{-c_3 x}, \quad (16)$$

where c_i are constants. I shall not consider this case any further, for there are strong reasons for supposing that a will be greater than one. First, $k(1+\sigma)/(1+k)$ is greater than one when $k\sigma$ is. We expect σ to exceed one by a lot, and k to be at least one. Next, any value of m short of infinity will push a even further away from one. Even if we take $k = 1$, for $\sigma = 3$ and $m = \infty$, we have $a = 2$, for $m = 1$, we have $a = 4$. For $\sigma = 9$, the corresponding values are five and ten. I shall proceed assuming $a > 1$, and then we have

$$z(x) = [c_1 + c_2 x]^{-1/(a-1)} \quad (17)$$

for the case of pure congestion costs.

The solution of (15) is now reduced to a quadrature, and an easy one if we transform to z as the independent variable. Then the numerical part of the problem is reduced to solution of algebraic equations corresponding to the boundary conditions. I shall return to numerical solutions in the fifth section.

■ The solution (17) leads to several interesting qualitative results. First, we can at once find the residential density as a function of the distance from the center. We have

$$-\frac{N'(x)}{G_1(x)} = \text{const} [c_1 + c_2 x]^{(1+[a/k(a-1)])}. \quad (18)$$

The constants depend on the rest of the solution, but the power depends only on the parameters. For example, if $k = 1$ and $m = \infty$, it is 3 for $\sigma = 3$ and 2.25 for $\sigma = 9$.

It is only in the special case $a = 1$ that the residential density function has a negative exponential form. Since this form has been extensively used, and defended by Mills and Muth,¹³ some explanation is called for. The main difference between my analysis and theirs is that they do not consider the road width as a policy variable, but fix it in advance. What is more, they assume a special form, $G_2(x) = g_2 x$ where g_2 is a constant between zero and one. This fixes the angle subtended by the roads at the center, so that the physical width of roads increases linearly away from the center. This is clearly a poor assumption. If anything, we should expect the road width to decrease away from the center in response to the lower traffic density. The equilibrium calculation is difficult for any other form of $G_2(x)$ laid down in advance. But if this function is chosen optimally, we get a very simple expression for the density. Even in practice, the assumption of optimality may be nearer the truth than $G_2(x) = g_2 x$, and therefore (18) should be worth using in estimation.

Another point of difference worth noting is that Mills and Muth do not study problems of congestion, and instead assume that $T'(x)$ is exogenous. Solow admits congestion, but in spite of assuming constant-angle roads, runs into the complexity of Bessel functions.¹⁴ It is interesting to note that when we allow $G_2(x)$ to be chosen optimally, following what Solow calls "an even more difficult alternative," we in fact simplify the problem greatly and find simple closed-form solutions that are natural generalizations of negative exponentials.

One generalization considered by Mills and Muth but neglected here is a more sophisticated treatment of housing. I have treated housing as being identical with land, instead of obtaining it by combining land and other inputs through a production function. A Cobb-Douglas production function for housing produces density profiles of the same functional form as (18); this is shown, for example, by Mills. He concludes that this source of departure from a negative exponential form is not important enough in practice,¹⁵ although he had argued in support of a form like (18) against a negative exponential.¹⁶ I hope that I have produced here much stronger reasons for the more general form.

Models of urban residential equilibrium usually hinge crucially around the rent function. It will therefore help us compare the two to find the implied rent function for the model of this paper.

¹³ See [2], chapter 5, and [6], chapter 4, respectively.

¹⁴ In [11].

¹⁵ In [3].

¹⁶ In [2].

We have

$$r(x) = \text{const} [c_1 + c_2 x]^{-[(1+k)a/k(a-1)]}. \quad (19)$$

In the Rawlsian case, the power is numerically equal to $(1 + \sigma) \times (1 + k)/(k\sigma - 1)$. For $k = 1$ and $\sigma = 3$, for example, it will be 4.

We can equally readily find the traffic density function:

$$\frac{N(x)}{G_2(x)} = \text{const} [c_1 + c_2 x]^{-[a/k(a-1)]}. \quad (20)$$

This, too, is negative exponential only when $a = 1$.

We can now combine (18) and (20) to obtain an equation for the optimum road pattern

$$G_2'(x) = \frac{ac_2}{a-1} \left[\frac{1+k}{k} G_2(x) - x \right] / (c_1 + c_2 x). \quad (21)$$

This can be solved in closed form by familiar methods, but the solution is not as instructive as simple inspection. Clearly, we have $G_2(s) = 0$, and therefore $G_2'(s) < 0$. Thus the physical width of roads will decrease away from the center at least near the outer boundary of the residential belt. We can now trace the solution backward. Of course $G_2(x)$ may equal x for some $\bar{x} > c$, and then all land between c and $\bar{x}$ will be used for traffic. This cannot occur when c is chosen optimally, for then we would put the load up to $\bar{a}$ to better use in production by increasing c . The three possibilities are shown in Figure 1. This will be considered further in the numerical section.

I shall consider next the distribution of utility. We have

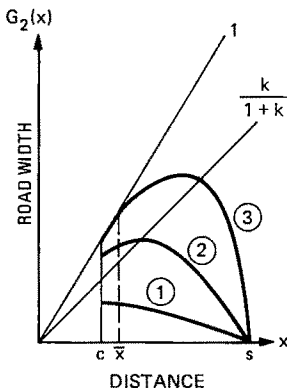
$$u(x) = \text{const} \cdot [c_1 + c_2 x]^{[1/m(a-1)]}. \quad (22)$$

For this model, it is possible to say unambiguously that households further away from the center get higher utility. They get fewer consumer goods, but enough extra space to turn the distribution of utility the other way.

It is easy to see that $m(a-1)$ increases as m increases, for we have $m(a-1) = k(1+\sigma)/(1+k) + m(k\sigma-1)/(1+k)$. Thus, as m increases, the distribution of utility becomes more even. In the limit, everyone has equal utility. The forms of additively separable welfare functions chosen in other contexts like optimal growth have substantially lower values of m . Something in the neighborhood of one is often thought to be reasonable. For such a value, the inequality involved can be substantial.

There is a major difference between the results here and those of Mirrlees.¹⁷ In his model, utility is equally distributed if preferences are homothetic between consumer goods and housing. In my model, preferences are homothetic but utility is unequally distributed. There are two ways in which my model differs from his. First, I consider congestion costs and he considers distance costs. Second, the costs appear on the production side of my model and on the consumption side of his. To be precise, he keeps labor hours fixed and deducts commuting time from leisure (and so from utility), while I keep leisure hours fixed and deduct commuting time from labor (and so

FIGURE 1
USE OF LAND FOR STREETS



from output).¹⁸ The role of these assumptions in determining the distribution of utility is studied by Riley and Stern.¹⁹

Finally, I shall mention some properties of the decentralized town, which corresponds to the optimum. General theorems of this nature have been proved by Mirrlees and Starrett. The model of this paper can be used as an illustration and as an aid to our understanding of the problem, because the special utility and production functions enable us to do simple aggregate calculations. The marginal product of labor is $\alpha Z/L$, so the wage bill is $W = \alpha Z$. The rest goes to the package of land and its complementary input, so land in the business district receives rent $R_b = (1 - \alpha)Z - \gamma E$. But the marginal product of the package is $\beta Z/E$, so the imputed rent is $R_b^* = \beta Z - \gamma E$. Because of increasing returns to scale, $R_b < R_b^*$.

Since each household spends $\sigma/(1 + \sigma)$ of its income on the consumer good and $1/(1 + \sigma)$ on rent, the total expenditures must be the same fractions of total household incomes. Let S be the net subsidy to households, i.e., total lump sum subsidies minus total receipts from congestion tolls. Then the total rent from the residential belt is $R_h = (W + S)/(1 + \sigma)$, and the condition for equilibrium in the output market is

$$Z - \gamma E - \pi \rho s^2 = \sigma(W + S)/(1 + \sigma).$$

This can be written as

$$R_b + R_h = \pi \rho s^2 + S. \quad (23)$$

If we suppose that the town government controls the use of its land, then (23) says that the optimum can be decentralized with the government budget balanced.

Put another way, we get a form of the theorem of Mirrlees and Starrett mentioned in the introduction. We have

$$\alpha + \beta - 1 = (R_b^* + R_h - \pi \rho s^2 - S)/Z. \quad (24)$$

Thus the degree of increasing returns equals the ratio of imputed differential land rents minus net household subsidy to the value of output. Note that for Mirrlees and Starrett, this is a *necessary* condition for the population to be optimum; here it holds for every population level provided the city is organized optimally for that level. This is a consequence of the various constant elasticity functions built into my model.

■ Throughout this section, population is measured in units of thousands of households (the population in thousands of people would be three to four times this), distance in miles (and area in square miles), and output in thousands of dollars per year. The value of the parameters is chosen in such a way as to make the model comparable to earlier work without losing its essential differences. For most of the time, I have taken $m = \infty$ to enable me to compare the model with equilibrium models. I have taken $\sigma = 3$ and $k = 1$, these being values as near as possible to those used in earlier work, without being too unrealistic. The base value of δ was

5. Numerical results

¹⁸ This was not intended to reflect any fundamental differences in priorities.

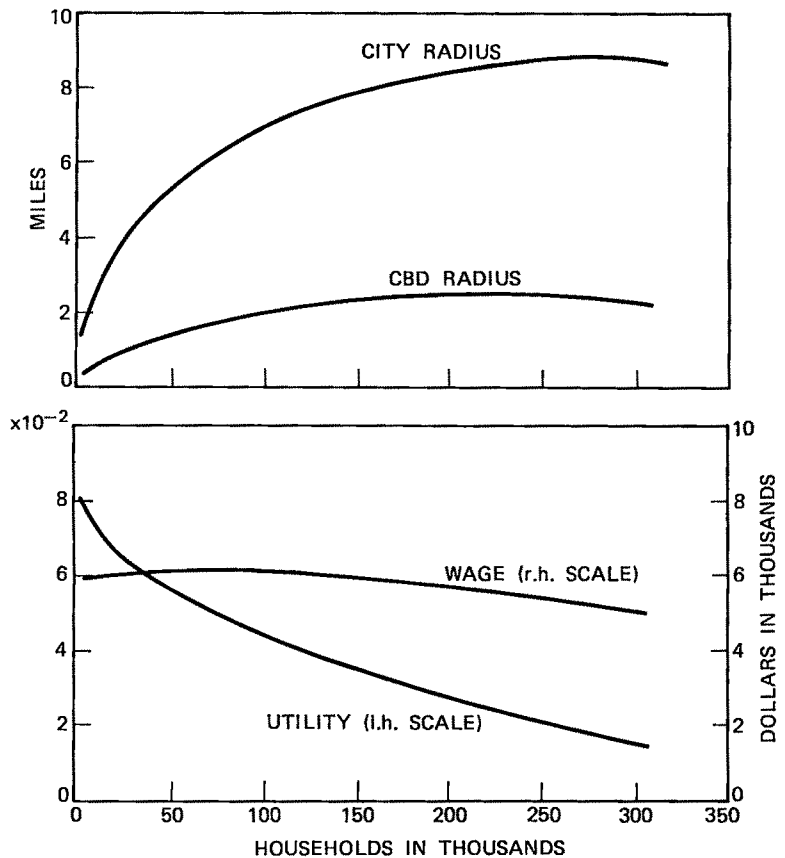
¹⁹ In [10] and [15], respectively.

chosen to be 10^{-6} , and some variations around it were studied. The base value had the property that in typical towns, the average household spent about 10 percent of its total nonleisure time in commuting, which seems reasonable. A very crude measure for γ was found from the average property taxes on industrial realty per unit of area, which accords with the second interpretation of γ but not with the first one mentioned earlier. This was deemed sufficient for the order-of-magnitude calculations in which I was interested. A casual inspection of Netzer²⁰ produced a value $\gamma = 15,000$. The rural rent ρ was taken to be 1500, as in some of Solow's cases.²¹ This value represents a reasonable ratio of the rural rent to the implied central business district rent. Some results of varying γ and ρ will be mentioned later.

I have kept α/β fixed at 2.5. Now for the Cobb-Douglas function, the ratio of labor to industrial land is α/β times the implied ratio of rents to wages. With my value of α/β , this turns out to be typically around fifty men per acre in my model, which is somewhat low. As a result, my towns turn out to be somewhat too large. I shall use this low value of α/β , because otherwise I could not study economies of scale to any significant extent without running into increasing returns to labor alone. The extent of economies of scale is

FIGURE 2

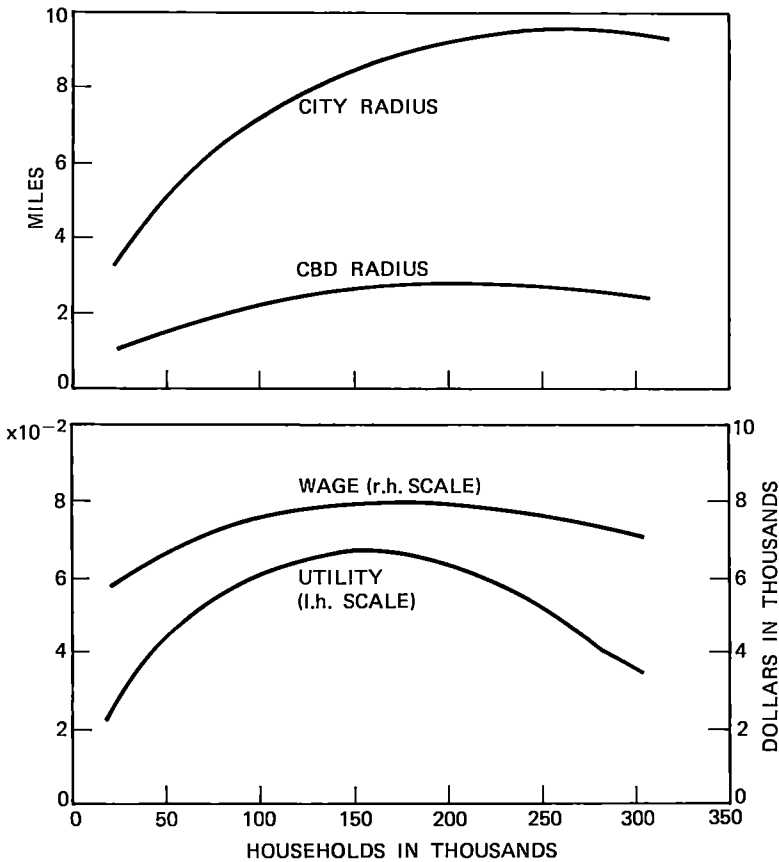
UTILITY AND WAGE RATE: $\alpha + \beta = 1.00$



²⁰ In [7].

²¹ See [13].

FIGURE 3
UTILITY AND WAGE RATE: $\alpha + \beta = 1.15$

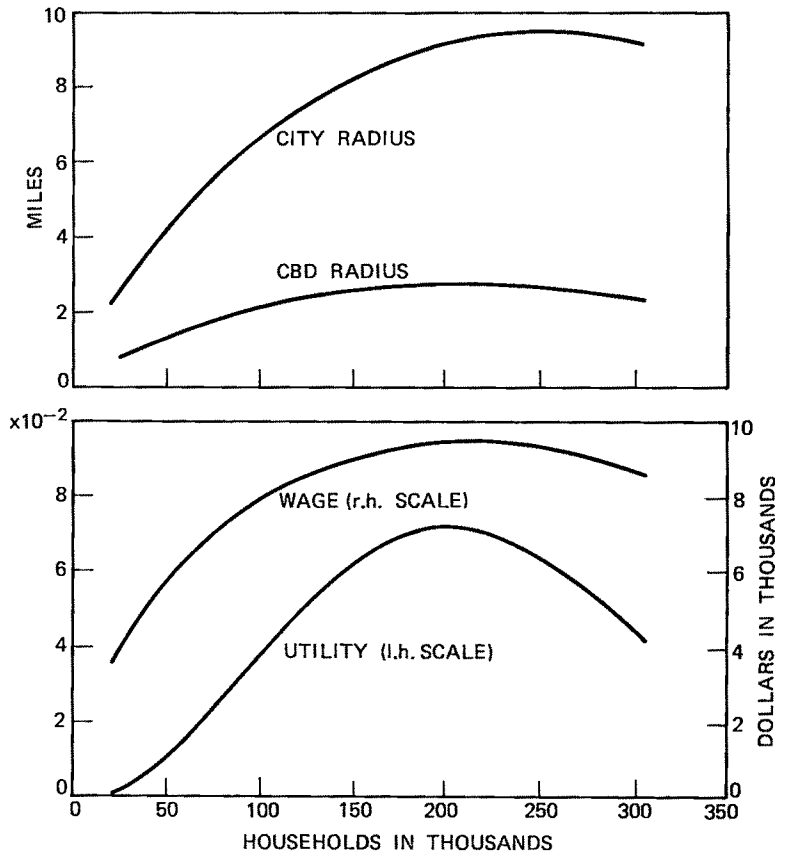


varied by changing α and β in the same proportion. This, I hope, minimizes any biased effect on the city size. At the same time, the multiplicative constant is changed to leave output the same for the previous optimum configuration. This leaves the old policy just feasible, and seems to be the best way in which to study the effect of changing economies of scale. The spectrum of such economies is scanned by chaining together small changes like this.

The first set of results (in Figures 2, 3, and 4) shows how the optimum arrangement of a town with a prescribed population depends on the level of this population, the remaining parameters being held fixed. Figure 2 has no economies of scale, Figure 3 has 15 percent and Figure 4 has 35 percent. In all cases, the utility of the representative household, the wage, the inner radius, and the outer radius are all unimodal functions of the population, and reach their maxima in the order mentioned. With no economies of scale, utility is maximum for zero population, showing dispersed production to be optimum.

We see this by answering the general problem posed at the beginning of Section 2, at least in the Rawlsian case. Write $u(N)$ for the utility of each household, when N of them are arranged optimally in a town. Now the utilitarian welfare index from the arrangement of the population P in optimal towns of N households each is $W(P, N) = (P/N) \cdot Nu(N) = Pu(N)$, which is maximum as a function of N

FIGURE 4
UTILITY AND WAGE RATE. $\alpha + \beta = 1.35$

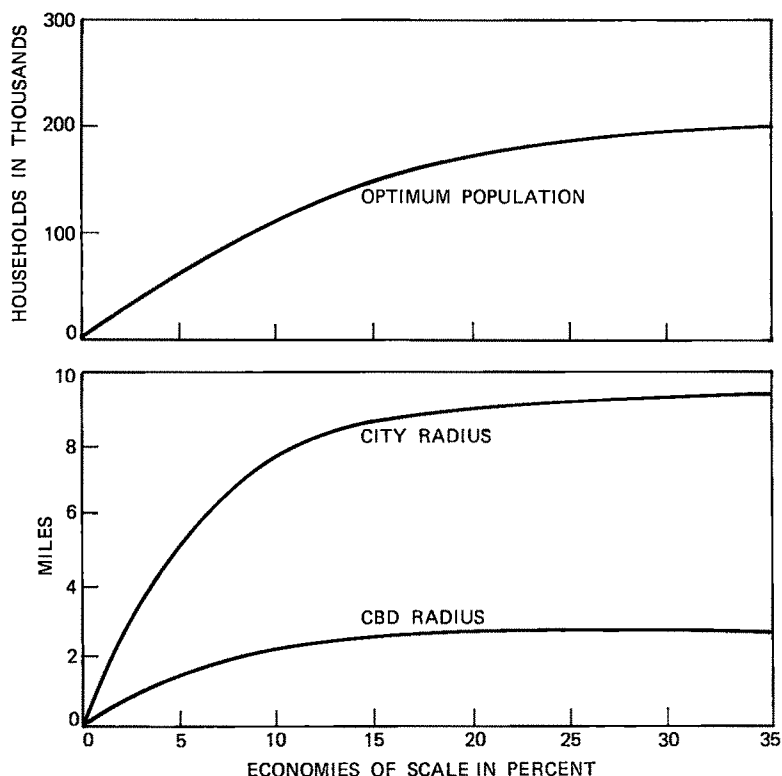


when $u(N)$ is. This simple argument is, of course, limited to the Rawlsian case; Mirrlees gives the general case.

Figures 2, 3, and 4 also show the magnitude of error that will be made if we neglect the congestion problem and use another index of welfare, such as the wage rate. The error is substantial for low levels of economies of scale. It is also interesting to note that the radii of the optimally arranged towns are not increasing functions of the population level, and begin to decrease for very congested towns. This is because commuting in these very congested towns takes a large amount of time, leaving little time for work and leading to low levels of output. The marginal utility of output then rises, and it is no longer worthwhile to pay rent for the larger area of land.

Figure 5 shows the overall welfare maximizing level of population of each town as a function of the economies of scale, and also shows the inner and outer radii of the corresponding optimally arranged towns. The first few percents of economies of scale lead to a rapid increase in the optimum population, but this effect wears off, and towns of about 200,000 households are the biggest we can get. Thus the model cannot explain Chicago as an optimum town. This is probably equally the fault of the model and of Chicago. The model does show, however, that the towns used in equilibrium calculations in similar models but without economies of scale are far too

FIGURE 5: OVERALL WELFARE MAXIMIZING LEVEL OF POPULATION



large to be optimum. Examples are found in the work of Solow and of Oron et al.²²

Variations in other parameters were studied next. It turned out that the optimum population had an elasticity of -1 with respect to δ , so that halving δ doubled the optimum population. This remarkably exact result must have a theoretical justification, but I have not been able to find one yet. It does show, however, that transport costs are very important in influencing optimum city size. Lower transport costs enable us to take greater advantage of economies of scale by having bigger cities, and the effect turns out to be numerically quite large.

The optimum population is quite insensitive to changes in A , γ , and ρ over very wide ranges. The optimum radii of the city depend on these in the obvious manner.

Next, numerical solutions to the differential equation (21) for optimum road width were obtained for several of the cases considered above. The solutions were always of the second type shown in Figure 1. The physical width of roads $G_2(x)$ increased for about a third to a half of the distance out into the residential belt (more for more congested cities) and then decreased to zero at its outer boundary. The proportion of area used for roads, $G_2(x)/x$, decreased everywhere. The proportion never reached one, even in cities so congested as to have a population twice the optimum. It did, however, rise as high as 75 percent near the edge of the central business district.

Finally, the valuation parameter m , held fixed at infinity up to now, was varied, and the extent of inequality in the optimum ar-

²² See [13] and [8], respectively.

rangement was studied. Given the crude nature of the whole numerical calculation, I did not think it necessary to use any measure more sophisticated than the ratio of the range to the average utility. The results are shown in Table 1. This illustrates how large the extent of

TABLE 1: EXTENT OF INEQUALITY IN OPTIMUM ARRANGEMENT

m	u_{MIN}	u_{MAX}	$200(u_{MAX}-u_{MIN})/(u_{MAX}+u_{MIN})$
1,000	0.3122	0.3123	0.03
100	0.3114	0.3128	0.45
10	0.3043	0.3171	4.12
5	0.2976	0.3212	7.63
2.5	0.2872	0.3276	13.14
1	0.2687	0.3392	23.19
0.25	0.2432	0.3554	37.49
PARAMETERS $\alpha + \beta = 1.15$, $N = 120,000$, $\delta = 10^{-6}$, $\gamma = 15,000$, $\rho = 1,500$.			

inequality can become for values of m that are thought to be reasonable in other contexts where additively separable welfare functions are used. As I mentioned earlier, this raises some doubts about the political feasibility of the inequality in the optimum town.

6. Concluding remarks

■ The limitations of this model need hardly any emphasis from me. Several improvements are possible even within the general framework. The most obvious is the introduction of transport costs within the central business district. These will imply a rent gradient within this district, and it will then become desirable to relax the assumption of fixed proportions between land and the composite input E , thereby allowing more intensive use of higher rent land. Housing services should be treated more realistically as a mixture of land and other inputs. The way in which more general transport cost functions can be allowed is clear from the theoretical work of Section 3, but numerical solution of the differential equation will then be necessary. Some or all of these aspects have been discussed by the authors mentioned in the introduction, and I have neglected these refinements in order to concentrate on the trade-off between economies of scale in production and congestion effects in transport.

There are important reasons why the whole framework falls short of the ideal for discussion of this problem. The nature of economies of scale is clearly more complex than that manifest in a production function with increasing returns for a homogeneous consumption good. A proper treatment must allow for multiple products and trade between towns. There are indivisibilities in transport, too, and analyzing these properly in relation to congestion needs dynamic analyses. Unfortunately, the ways to make the framework used here dynamic are not very clear.

I acknowledge the force of all these difficulties. Nevertheless, the functional forms and numerical results obtained here seem reasonable enough to be workable first approximations, and I think that some progress towards improving our understanding of urban problems has been made.

References

1. KOOPMANS, T. C. *Three Essays on the State of Economic Science*. New York: McGraw-Hill, 1957.
2. MILLS, E. S. "An Aggregate Model of Resource Allocation in a Metropolitan Area." *The American Economic Review*, Vol. 57, No. 2 (May 1967), pp. 197-210.
3. ———. *Urban Economics*. Oakland, N. J.: Scott Foresman, 1972.
4. ——— AND DE FERRANTI, D. M. "Market Choices and Optimum City Size." *The American Economic Review*, Vol. 61, No. 2 (May 1971), pp. 340-345.
5. MIRRELES, J. A. "The Optimum Town." *Swedish Journal of Economics*, Vol. 74, No. 1 (March 1972), pp. 114-135.
6. MUTH, R. *Cities and Housing*. Chicago: University of Chicago Press, 1969.
7. NETZER, D. *Economics of the Property Tax*. Washington, D. C.: Brookings Institution, 1966.
8. ORON, Y., PINES, D. AND SHESHINSKI, E. "Optimum versus Equilibrium Land Use Patterns and Congestion Toll." *The Bell Journal of Economics and Management Science*, Vol. 4, No. 2 (Autumn 1973), pp. 619-636.
9. RAWLS, J. *A Theory of Justice*. Cambridge: Harvard Univ. Press, 1971.
10. RILEY, J. G. "Optimal Towns." Unpublished Ph.D. dissertation, M.I.T., 1972.
11. SOLOW, R. M. "Congestion, Density and the Use of Land in Transportation." *Swedish Journal of Economics*, Vol. 74, No. 1 (March 1972), pp. 161-173.
12. ———. "On Equilibrium Models of Urban Location," in J. M. Parkin, ed., *Essays in Modern Economics*, Longmans, 1973.
13. ———. "Congestion Cost and the Use of Land for Streets." *The Bell Journal of Economics and Management Science*, Vol. 4, No. 2 (Autumn 1973), pp. 602-618.
14. STARRETT, D. A. "On the Optimal Degree of Increasing Returns," Discussion Paper No. 230, Harvard Institute of Economic Research, 1972.
15. STERN, N. H., "Homogeneous Utility Functions and Equality in 'The Optimum Town.'" *Swedish Journal of Economics*, Vol. 75, No. 2 (June 1973).
16. VICKREY, W. "Pricing as a Tool in Coordination of Local Transportation," in *Transportation Economics*, National Bureau of Economic Research, 1965.

SHORT ARTICLES, COMMENTS AND REVIEWS

- 655 A further analysis of provincial trucking regulation
J. PALMER
- 665 Investment behavior in the U. S. telephone industry—1949
to 1968
U. SANKAR
- 679 Ralph Turvey's Economic Analysis and Public Enterprises
W. G. SHEPHERD
- 682 George C. Eads' The Local Service Air Line Experiment
P. VERLEGER
- 685 The FTC staff's Economic Report on Conglomerate Merger
Performance
J. F. WESTON
- 690 Leonard Waverman's Natural Gas and National Policy: A
Linear Programming Model of North American Natural Gas
Flows
D. KENDRICK
- 693 Jan Mossin's Theory of Financial Markets
M. E. RUBINSTEIN
- 700 Omissio

A further analysis of provincial trucking regulation

John Palmer

Assistant Professor of Economics
The University of Western Ontario

Forcing what is essentially a hyperbolic relationship into a linear regression model can create a "hyperbolic bag." This bag takes on importance if it is systematically related to other variables in the regression model. For the case at hand, it appeared that using a linear model might overstate the effects of regulation on trucking rates because the provinces which regulate rates have either relatively short or relatively long average distances, while those which do not regulate rates have intermediate distances falling at the bottom of such a bag. The regression model was respecified to correct for this problem, and while the respecification appears unimportant, other questions are raised.

■ "When I had scarcely learned to toddle,
My parents handed me a model
With lots of little leads and lags . . .
And pretty [hyperbolic] bags."
—paraphrased from D. H. Robertson.

In the past few years two different but similar articles by Sloss and McLachlan¹ have appeared which use cross-sectional analysis to test the effects of different rate-regulatory schemes in different provinces on the average revenue per ton-mile in intraprovincial Canadian trucking. This paper takes those authors to task for using econometric models which imply incorrect economic theory, and while the results of the tests reported here are in many ways similar to theirs, it appears that perhaps less definitive statements can be made than have previously been made.

■ Sloss initially studied the trucking regulations of the provinces for the period 1958 to 1963 and divided them into the following two categories:

<u>Regulated:</u>	<u>Nonregulated:</u>
British Columbia	Alberta
Saskatchewan	Ontario
Manitoba	New Brunswick
Quebec	Nova Scotia
	Prince Edward Island
	Newfoundland.

Financial assistance was provided for this study by the Canadian Consumer Council, though the Council is in no way responsible for the contents of this paper. The author wishes to thank John Erkkila for valuable research assistance, and James Sloss for making available some of his unpublished results.

¹ In [2] and [1], respectively.

1. Introduction

2. Sloss' study

He then performed residual analysis on the following regression:

$$\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + \mu, \quad (1)$$

where

$\hat{Y}$ = Estimated average revenue/ton-mile for intraprovincial trucking for each year (1958 to 1963) for each province,

X_1 = Average length of haul for each year in each province,

X_2 = Average net weight per loaded vehicle,

X_3 = Average fuel tax/gallon,

X_4 = Average license costs per truck or tractor/year, and

X_5 = Average annual wage/employee.

Sloss next examined the residuals ($Y - \hat{Y}$) to determine whether they were systematically related to whether the provinces regulated trucking rates. The results must have been inconclusive, for he declared, "The experimental trials portended generally successful results, subject, however, to certain modifications."² These modifications included, *inter alia*, deleting Prince Edward Island and Newfoundland from his sample and changing Quebec from the regulated category to the nonregulated category.

His *ex post* justifications for the deletions were³:

Newfoundland and Prince Edward Island were excluded from the model because of (1) the relatively small size of their for-hire truck populations, (2) their insular locations, resulting in unique transportation problems, (3) caution expressed in D.B.S. reports that estimates for these provinces contained a large degree of sampling error, and (4) the impracticability of prorating to these provinces a reasonably accurate share of their costs and revenues when published in consolidated form for the Atlantic Provinces as a whole.

His reason for changing the classification of Quebec, aside from the fact that "initial quantitative analyses produced results for this province which differed markedly from *a priori* expectations and from results calculated for the other provinces which had been classified 'Regulating' "⁴ was that while rates are *de jure* regulated in Quebec, they are *de facto* not regulated.

These modifications yielded the following coefficients (and standard errors):

$$\begin{aligned} \hat{Y} = 9.7691 - 0.0102X_1 - 0.7779X_2 + 0.1983X_3 - 0.0096X_4 \\ (0.0056) \quad (0.1807) \quad (0.0740) \quad (0.0029) \\ + 0.0021X_5, \quad (R^2 = 0.7241). \quad (2) \\ (0.0007) \end{aligned}$$

The residual analysis revealed that the regulated provinces had significantly higher rates than the unregulated provinces and that regulation increased the average revenue/ton-mile by about $\frac{7}{8}$ of a cent/ton-mile.

Sloss should probably not be criticized too harshly for his modifications of the data. Although they were made *ex post*, a generally dangerous practice, they seem reasonable. One usually prefers to avoid nonrandom disturbances in the data which are unaccounted for by the independent variables, and such appears to be the case

² [2], p. 358.

³ *Ibid.*

⁴ *Ibid.*, p. 340.

for Prince Edward Island and Newfoundland. Also, if one is interested in the effects of *de facto* regulation, Sloss makes a strong case for classifying Quebec as nonregulated.

One minor irritant remains, however—the inexplicably significant negative coefficient for average license costs (X_4). If license costs have any effect on trucking rates, they would generally be expected to have a positive effect on them, but Sloss ignores this puzzling result.

McLachlan tested the same hypothesis using the same dependent variable over the same time period, but with a different classificatory scheme, a different sample, different statistical techniques, and different independent variables. First, he left Quebec classified as regulated and by default tested the effects of *de jure* regulation. Second, using data from only those provinces which were more or less clearly regulated or nonregulated, he was left with:

3. McLachlan's study

<u>Regulated:</u>	<u>Nonregulated:</u>
British Columbia	Alberta.
Saskatchewan	
Manitoba	
Quebec	

Rather than employ residual analysis, McLachlan included a dummy variable for regulation in his regression. The preliminary equation fitted was

$$Y = K + a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 + a_5X_5 + a_6X_6 + a_7X_7 + a_8X_8 + \mu, \tag{3}$$

where

- Y = Average revenue/ton-mile,
- X_1 = Average wage rate/hour in each province,
- X_2 = Average cost/gallon of fuel, inclusive of tax, in each provincial capital,
- X_3 = An index of the relative costs of transport in each provincial capital,
- X_4 = An allowance for the passage of time,
- X_5 = A regulation dummy variable: $X_5 = 1$ if regulated, zero otherwise,
- X_6 = Average annual license costs,
- X_7 = Average miles that each ton was carried, and
- X_8 = Average weight.

McLachlan initially ran a regression including all the independent variables, but then reran it omitting those which were statistically insignificant in the initial run. The results he reported were:

$$Y = 0.68383 + 1.61390X_1 + 0.08007X_3 + 2.58190X_5 - 0.03697X_7 - 0.34679X_8 \tag{4}$$

(0.64615)
(0.02080)
(0.27699)
(0.00357)
(0.11767)

(standard errors in parentheses, $R^2 = 0.96$).

These results are similar to Sloss' in that regulation appears to have a significantly positive effect on trucking rates, but in this case it contributes about $2\frac{1}{2}$ cents/ton-mile to trucking rates. Average license costs, with their surprisingly significant negative coefficient in Sloss' equation, proved insignificant for McLachlan.

4. The error of the linear regression model

■ Both Sloss and McLachlan committed a potentially serious error in using a linear regression model to estimate what is essentially a *nonlinear* cost-determined rate structure, particularly with respect to the distance variable.

To illustrate this point, suppose that all factors other than distance which affect costs are held constant. Then a reasonable revenue function (assuming revenues are cost-determined) would be of the form

$$R_i = a + bd_i, \quad (5)$$

where R_i is the revenue derived from the i th trip, a is a constant accounting for all costs which do not vary with distance, d_i is the distance of the i th trip, and b is the marginal cost of travelling an additional mile. This marginal cost is assumed to be approximately constant, accounting for depreciation, fuel consumption, etc.⁵

Given the above equation, we can derive the appropriate form of the statistical model to study $\frac{\sum R_i}{\sum d_i}$, a form of the dependent variable used by both authors:

$$\begin{aligned} \sum R_i &= na + b \sum d_i \\ \frac{\sum R_i}{\sum d_i} &= \frac{na}{\sum d_i} + b \end{aligned} \quad (6)$$

or

$$\frac{\sum R_i}{\sum d_i} = \frac{a}{\bar{d}} + b.$$

This derivation implies a hyperbolic average cost curve, with respect to distance, asymptotically approaching b as $\bar{d}$ increases. By forcing the data into a linear model, both authors have created a hyperbolic bag below their regression line (see Figure 1).

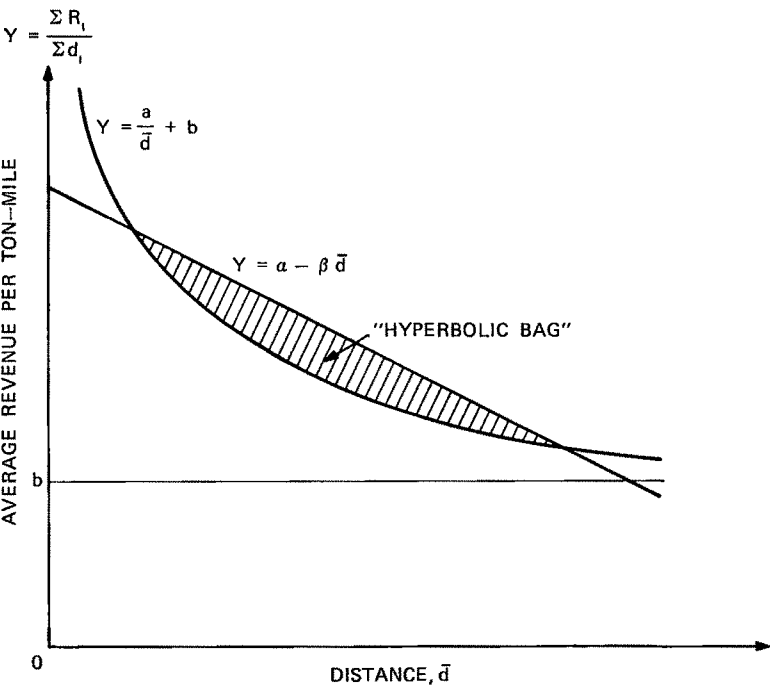
5. The hyperbolic bag

■ The creation of this bag seems relatively unimportant *unless* those provinces which were unregulated happened to have $\bar{d}$'s which fell in the bag (generating negative residuals for Sloss), while those which were regulated had $\bar{d}$'s which fell outside the bag and above the regression line (generating positive residuals). If this hyperbolic bag exists, one would expect the residuals of Sloss' equations when plotted against $\bar{d}$ to appear as in Figure 2. To examine this prediction, the residuals were plotted in Figure 3. While somewhat scattered, they do conform in a general fashion to the prediction of Figure 2.

What is more important, however, is that the aforementioned strange coincidence seems to have existed! The unregulated provinces,

⁵ This formulation of the model is suggested by auto and truck rental rates, usually of the form \$X/day, \$Y/mile.

FIGURE 1
HYPOTHESIZED "HYPERBOLIC BAG"



most notably Alberta, New Brunswick, and Quebec, fell into the bag, while the regulated provinces, particularly British Columbia, Saskatchewan, and (to some extent) Manitoba fell at the extremes, outside the bag.

One might reasonably expect this bag not to exist for the weight variable, since the marginal costs of carrying an additional ton are probably not constant. A plot of Sloss' residuals against weight (Figure 4) revealed no striking pattern and confirmed this expectation.

But spurred on by this discovery concerning the distance hyperbolic bag, I then proceeded to run 24 different regressions to see

FIGURE 2
RESIDUALS EXPECTED IF HYPERBOLIC BAG EXISTS

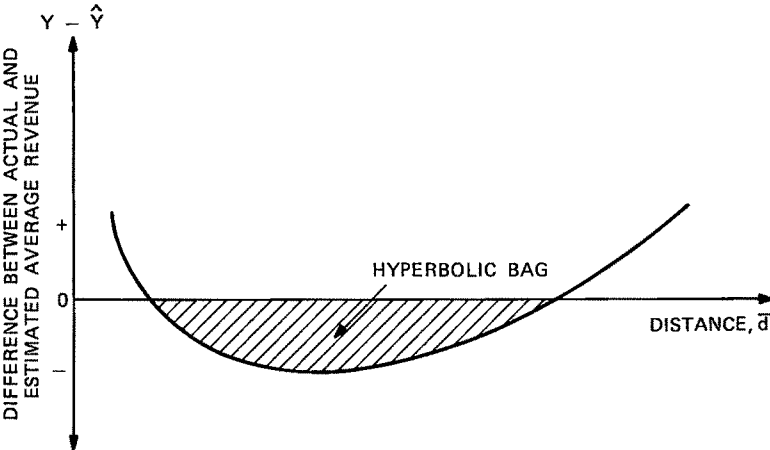
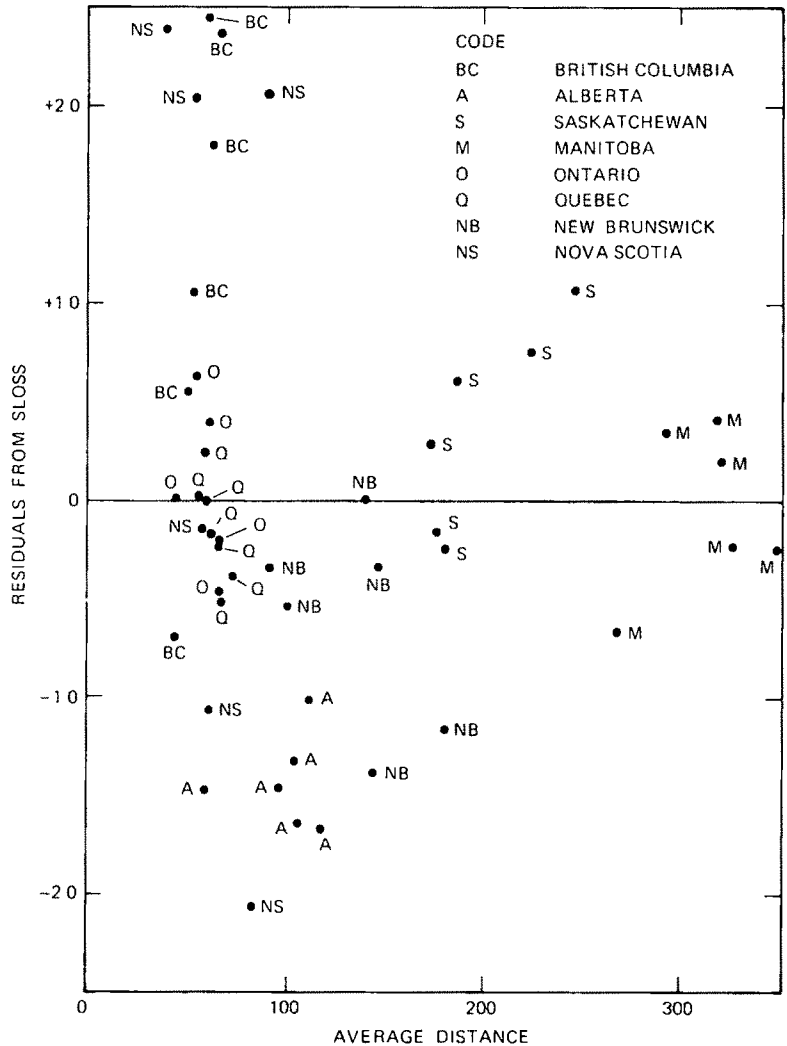


FIGURE 3
RESIDUALS PLOTTED AGAINST AVERAGE DISTANCE

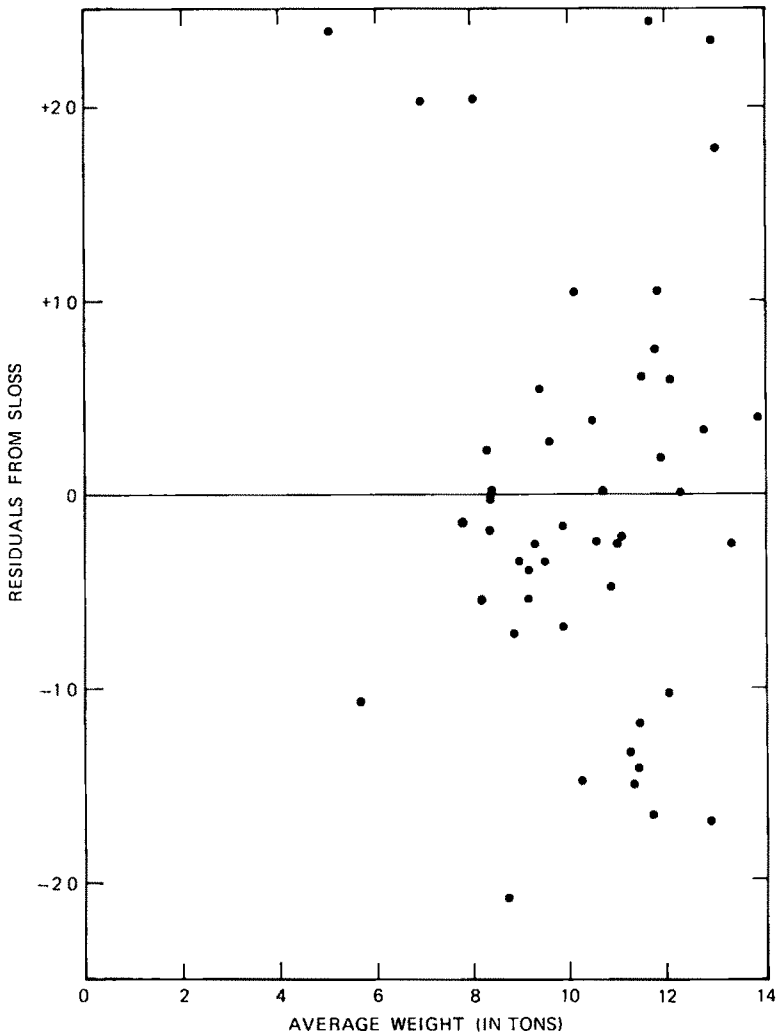


whether using $1/\bar{d}$ might change the results of Sloss and McLachlan concerning the effects of regulation.

Three different samples of provinces were used. The first included all provinces, with the recognition that Newfoundland and Prince Edward Island statistics were unreliable and might be affected by other nonincluded independent variables peculiar to their locations; I included these provinces in one set of regressions, though, to see whether the changes I made in the model would affect the results when they were included. The second sample was the same as Sloss': all provinces except Newfoundland and Prince Edward Island. The third sample was that of McLachlan: British Columbia, Alberta, Saskatchewan, Manitoba, and Quebec.

Eight different regressions were run for each sample, to allow the following variations. In half of the regressions $1/\text{weight}$ was used as an independent variable rather than *weight*. In half of the regressions, Quebec was classified as regulated, in half unregulated. And in half of them average miles/gallon of fuel was included as an additional independent variable in the hope that it might proxy for differences in terrain and pick up some additional costs. It was omitted from the

FIGURE 4
RESIDUALS PLOTTED AGAINST AVERAGE WEIGHT



other half because it was expected *a priori* to have a strong inverse relationship to weight.

■ Tables 1 and 2 present the results of these regressions along with an explanation of the variables used. The most surprising result was that in 18 of the 24 regressions the dummy variable for regulation was significant! I certainly did not expect this result, based on my findings concerning the distance variable and Sloss' residuals. And the six regressions in which the regulation dummy was not significant all were run on the first sample, which included Newfoundland and Prince Edward Island. One possible explanation for these exceptions is that in spite of the fact that trucking rates in these two provinces were unregulated, truckers there had high fixed costs due to the problems of transporting new trucks there; and one must also keep in mind Sloss' caution that data from these provinces may be unreliable.

The coefficient of $1/d$ had the expected sign and was significant in all but one regression. *Weight* and $1/weight$ also had the expected

6. Regression results

TABLE 1

REGRESSION RESULTS WITH MILES/GALLON EXCLUDED

REGRESSION	INDEPENDENT VARIABLES										
	CONSTANT	1/DIST.	WT.	1/WT.	TIME	FUEL TAX	LIC. FEE	WAGE	REG.	REG.**	R ²
SAMPLE I	20.19	118.72* (9.34)	-0.781* (-3.80)		0.613* (2.59)	0.00015 (0.00)	-0.0054 (-1.30)	-0.0789 (-1.22)	1.79 (1.85)		0.914
	22.07	132.27* (8.20)		17.674 (1.77)	0.741* (2.89)	-0.067 (-0.63)	-0.011* (-2.55)	-0.191* (-3.09)	2.210* (2.10)		0.896
	14.32	117.24* (8.76)	-0.832* (-3.92)		0.472* (2.05)	0.022 (0.21)	-0.0002 (-0.06)	-0.026 (-0.42)		0.329 (0.48)	0.909
	14.94	131.15* (7.84)		19.81 (1.94)	0.576* (2.30)	-0.03 (-0.27)	-0.0049 (-1.67)	-0.137* (-2.45)		0.753 (1.03)	0.890
SAMPLE II	20.24	219.52* (7.35)	-0.309* (-3.18)		0.444* (4.37)	0.0041 (0.11)	-0.0048* (-2.51)	-0.163* (-5.40)	1.890* (5.48)		0.938
	14.49	176.32* (5.19)		30.05* (3.97)	0.388* (3.93)	0.011 (0.30)	-0.0068* (-3.75)	-0.147* (-5.03)	2.042* (6.34)		0.945
	11.49	275.95* (7.89)	-0.250* (-2.09)		0.247* (2.19)	0.073 (1.65)	0.0022 (1.67)	-0.120* (-3.67)		0.868* (3.37)	0.916
	6.51	247.46* (6.16)		23.29* (2.49)	0.191 (1.66)	0.082 (1.87)	0.0011 (0.76)	-0.107* (-3.18)		0.932* (3.87)	0.919
SAMPLE III	13.53	69.89* (2.12)	-0.373* (-5.46)		0.026 (0.22)	0.081 (1.45)	-0.0094* (-6.27)	-0.010 (-0.31)	2.129* (9.71)		0.979
	3.66	59.23 (1.71)		43.52* (5.31)	-0.063 (-0.48)	0.092 (1.61)	-0.0090* (-5.92)	0.012 (0.34)	2.066* (9.21)		0.979
	0.57	93.39* (3.11)	-0.307* (-4.73)		-0.255* (-2.34)	0.067 (1.27)	-0.0026* (-2.71)	0.102* (3.35)		1.786* (10.50)	0.982
	-7.43	82.32* (2.64)		36.50* (4.76)	-0.325* (-2.84)	0.077 (1.44)	-0.0025* (-2.61)	0.120* (3.84)		1.748* (10.18)	0.982
DEPENDENT VARIABLE = AVERAGE REVENUE/TON-MILE FOR EACH PROVINCE FOR EACH YEAR, 1958 TO 1963. 1/DIST. = THE RECIPROCAL OF THE AVERAGE DISTANCE EACH TON WAS CARRIED. WT. = THE AVERAGE NUMBER OF TONS CARRIED PER TRIP. 1/WT. = THE RECIPROCAL OF WT. MILES/GAL = THE AVERAGE MILES/GALLON OF GASOLINE EACH YEAR IN EACH PROVINCE. TIME = A TREND EQUAL TO 0 IN 1958, 5 IN 1963. FUEL TAX = THE AVERAGE FUEL TAX/GALLON OF GASOLINE EACH YEAR IN EACH PROVINCE. LIC. FEE = THE ANNUAL COST OF A LICENSE FOR A DIESEL TRACTOR SEMI-COMBINATION EACH YEAR IN EACH PROVINCE. WAGE = THE AVERAGE WAGE RATE/WEEK EACH YEAR IN EACH PROVINCE. REG. = 1 IF THE PROVINCE REGULATED TRUCKING RATES, 0 OTHERWISE (ASSUMING QUEBEC UNREGULATED). REG.** = REG., BUT ASSUMING QUEBEC REGULATED.											
t-VALUES IN PARENTHESES, * INDICATES STATISTICAL SIGNIFICANCE AT THE 5-PERCENT LEVEL											

signs in each regression, but *weight* proved to be significant more often than its inverse.

The time trend yielded puzzling results. One might expect this variable to have picked up such influences as the rate of inflation, the rate of technological change, or regional shifts in demand. Yet in general, when more provinces were included, the time trend had a significantly positive coefficient, while when fewer provinces were included, it had a negative (and sometimes significantly so) coefficient. A possible explanation of this result might be that the rates of inflation and innovation had differential effects on different provinces, depending on the relative growth patterns of the trucking industry in each province.

The coefficient of the fuel tax variable was usually positive, as expected, but significant in only two of the regressions.

The coefficient of the license fee variable seems to defy explana-

TABLE 2
REGRESSION RESULTS WITH MILES/GALLON INCLUDED¹

REGRES- SION	INDEPENDENT VARIABLES ²											
	CON- STANT	1/DIST.	WT.	1/WT	MILES/ GAL	TIME	FUEL TAX	LIC FEE	WAGE	REG.	REG.**	R ²
SAMPLE I	29.03	124.82* (9.11)	-0.902* (-3.95)		-0.686 (-1.17)	0.714* (2.84)	-0.055 (-0.51)	-0.0068 (-1.57)	-0.108 (-1.56)	1.86 (1.93)		0.915
	30.46	132.29* (8.20)		28.95 (1.95)	-0.881 (-1.02)	0.875* (3.04)	-0.140 (-1.10)	-0.0128* (-2.74)	-0.221* (-3.25)	2.27* (2.15)		0.893
	21.68	122.26* (8.53)	-0.942* (-3.89)		-0.594 (-0.98)	0.553* (2.26)	-0.026 (-0.02)	-0.0011 (-0.34)	-0.047 (-0.73)		0.297 (0.44)	0.910
	21.91	131.03* (7.82)		29.57 (1.92)	-0.755 (-0.85)	0.687* (2.43)	-0.093 (-0.69)	-0.0065 (-1.86)	-0.161* (-2.57)		0.741 (1.01)	0.891
SAMPLE II	14.13	203.83* (7.06)	-0.232* (-2.39)		0.624* (2.44)	0.388* (3.94)	0.041 (1.06)	-0.0044* (-2.43)	-0.149* (-5.11)	1.94* (5.96)		0.946
	11.99	180.70* (5.37)		22.91* (2.55)	0.410 (1.42)	0.369* (3.74)	0.031 (0.81)	-0.0061* (-3.29)	-0.146* (-5.07)	2.05* (6.46)		0.947
	6.28	263.63* (7.52)	-0.197 (-1.55)		0.510 (1.65)	0.197 (1.72)	0.104* (2.20)	0.0028 (1.92)	-0.107* (-3.21)		0.862* (3.42)	0.921
	4.31	251.41* (6.22)		17.18 (1.54)	0.352 (0.99)	0.175 (1.51)	0.099* (2.11)	0.0017 (1.10)	-0.106* (-3.16)		0.930* (3.86)	0.921
SAMPLE III	9.71	84.09* (2.79)	-0.313* (-4.75)		0.545* (2.51)	0.089 (0.80)	0.080 (1.59)	-0.0080* (-5.55)	-0.027 (-0.88)	2.20* (11.06)		0.984
	1.94	76.33* (2.31)		36.03* (4.32)	0.503* (2.16)	0.016 (0.12)	0.088 (1.65)	-0.0078* (-5.19)	-0.010 (-0.27)	2.15* (10.17)		0.982
	-3.25	107.36* (3.84)	-0.252* (-3.97)		0.486* (2.36)	-0.208* (-2.05)	0.067 (1.39)	-0.0012 (-1.09)	0.091* (3.22)		1.831* (11.73)	0.986
	-9.38	98.85* (3.28)		29.67* (3.77)	0.449* (2.09)	-0.265* (-2.40)	0.073 (1.48)	-0.0012 (-1.09)	0.104* (3.28)		1.800* (11.13)	0.985
¹ t-VALUES IN PARENTHESES, * INDICATES STATISTICAL SIGNIFICANCE AT THE 5-PERCENT LEVEL												
² SEE TABLE 1 FOR EXPLANATION OF VARIABLES												

tion. It is negative in twenty of the twenty-four regressions and significantly negative in twelve of the regressions. It seems implausible that, *ceteris paribus*, provinces with higher license fees should have lower rates, yet this appears to be the indication of these and Sloss' results.⁶

Similarly, the coefficient for the provincial wage indices yields puzzling results. That it should be negative in so many cases is surprising, but that it is significantly negative is even more surprising. It is possible, though, especially in light of its positive significance for the smallest sample, that from 1958 to 1963 there may have been regional shifts in the demand for trucking services in such a manner that revenues were not entirely cost-determined, and hence the wage rates and trucking rates were not so closely linked during the adjustment process.

In spite of these problems, the *R*²'s are all high, and the coefficients of the regulation dummies are strongly significant. It appears that one might still safely conclude that *de facto* regulation generally leads to rates which are about two cents higher per ton-mile while *de jure* regulation increases rates by somewhere between 0.9 cents and 1.8 cents. However, the problems still exist; if there were regional shifts in demand and different regional rates of inflation from 1958 to 1963,

⁶ Professor Sloss has suggested to me that perhaps this negative sign is the result of provinces with lower trucking rates using larger trucks with and in spite of the attendant higher license fees.

and if the trucking industry did not respond fairly rapidly to these regional differences, one suspects that trucking rates during that period were *not* entirely cost-determined, and these results, along with previous ones, are open to question pending further research.

References

1. McLACHLAN, D. L. "Canadian Trucking Regulations." *The Logistics and Transportation Review*, Vol. 8, No. 1 (1972), pp. 59–81.
2. SLOSS, J. "Regulation of Motor Freight Transportation: A Quantitative Evaluation of Policy." *The Bell Journal of Economics and Management Science*, Vol. 1, No. 2 (Autumn 1970), pp. 327–366.

Investment behavior in the U. S. telephone industry—1949 to 1968

U. Sankar

Associate Professor of Economics
University of Wisconsin—Milwaukee

This paper applies the econometric model of investment behavior developed in our earlier paper to the U. S. telephone industry. Utilizing annual data for the period 1949 to 1968, the author finds that the estimates of long-run elasticities of capital stock with respect to relative price and output are 0.51119 and 1.18225, respectively. However, the Cobb-Douglas specification of Jorgenson and Handel appears to be valid for this industry. In comparing the results for the telephone industry with those for the electric utility industry, substantial differences in technological characteristics and in the time form of lagged response of net investment are found.

■ In a previous paper¹ the author generalized the Jorgenson-Handel model of investment behavior² and applied it to the U. S. electric utility industry for the period 1949 to 1968. He found that their assumptions concerning the production function are not valid for that industry. The objectives of this paper are (1) to apply the generalized model to the U. S. telephone industry for the period 1949 to 1968, and (2) to test hypotheses pertaining to the long-run elasticities of capital with respect to output and relative price.

The plan of this paper is as follows. Section 2 reviews and compares the econometric models of Clark, Hickman, and Jorgenson and Handel.³ Section 3 summarizes the model developed in the previous paper⁴ and compares it with other models. Section 4 deals with data sources and measurement of variables. In Section 5 results for the telephone industry are presented and compared with those obtained for the electric utility industry. Section 6 contains some concluding remarks.

■ **Clark's model.** Clark's capital-requirements theory of investment⁵ modifies the pure acceleration principle. His theory may be summarized in three hypotheses.

The first hypothesis is that the firm's demand estimates depend via a fixed expectations coefficient upon the recent trend of its output, i.e.,

$$Q_{t+1}^e = Q_t \{1 + \epsilon[(Q_t - Q_{t-1})/Q_{t-1}]\},$$

The author wishes to thank the American Telephone and Telegraph Company for financial support.

¹ [8].

² See [5].

³ See [1], [3], and [5], respectively.

⁴ [8].

⁵ In [1].

1. Introduction

2. Review of econometric studies of investment behavior in the telephone industry

where Q_{t+1}^e is the expected output in period $t + 1$, Q_t is output in period t , and the expectations coefficient ϵ is defined as

$$\epsilon = \frac{Q_{t+1}^e - Q_t}{Q_t} \bigg/ \frac{Q_t - Q_{t-1}}{Q_{t-1}}.$$

If ϵ is zero, future output is expected to be the same as current output. If ϵ is one, the current proportional trend is expected to continue. The length of the period is the time required for construction. This implies a fixed lag of one period.

The second hypothesis is that the firm's net investment depends via a fixed capital coefficient and a fixed spare-capacity coefficient upon the demand estimates. This hypothesis can be written as

$$K_t = \beta C_t$$

$$C_t = \gamma Q_t^e, \quad \gamma > 1$$

$$\Delta K_t = K_{t+1} - K_t = \beta \gamma Q_{t+1}^e - K_t,$$

in which K_t is the firm's stock of capital equipment at the end of the period, C_t is the capacity of the stock of capital equipment defined in the same unit as output, β is the fixed technological ratio which the firm is assumed to maintain between its output and its stock of capital equipment, and γ is the spare-capacity coefficient.

The third hypothesis is that the firm's retirement depends via a fixed coefficient upon its present stock of capital equipment. This hypothesis may be written as

$$R_t = \rho K_t,$$

where ρ is the retirement coefficient.

Since he assumes fixed factor proportions, changes in the rate of interest and in the prices of output, of capital, and of other inputs have no effect upon the firm's investment. He recognizes that the coefficients β , γ , and ρ may change over time and interprets these changes as shifts in the technology.

Utilizing annual data relating to the telephone industry for the period 1912 to 1948, he obtains estimates of these coefficients for each year. His results indicate that these coefficients are not stable over time.

□ **Hickman's model.** Hickman's model for the telephone industry⁶ has three components: (1) demand for desired capital stock, (2) demand for replacement investment, and (3) the relationship between changes in demand for capital and actual changes in capital.

He assumes that the desired stock of fixed plant and equipment depends on expected or normal output, expected relative factor price, and time. His expression for desired capital stock is

$$K_t^+ = A(Q_t^e)^a [(w/c)_t^e]^b \exp(ct),$$

where K_t^+ is the desired capital stock in year t , Q_t^e is the output in year t , $(w/c)_t^e$ is the expected relative factor price (wage-rental ratio) in year t , and t is a time trend. He assumes that replacement investment is proportional to net capital stock. He assumes that the relationship between changes in desired capital stock and actual

changes in capital stock is of the following form:

$$(K_t/K_{t-1}) = (K_t^+/K_{t-1})^\beta \quad 0 < \beta \leq 1.$$

This is a logarithmic form of the flexible accelerator.

He estimates the following regression function:

$$\ln(K_t/K_{t-1})$$

$$= \beta[\ln A + a \ln Q_t^e + b \ln (w/c)_t^e + ct - \ln K_{t-1}] + U_t,$$

where U is a random error term. He assumes that expected output (and relative factor price) is a weighted average of current and recent outputs (factor prices). Utilizing annual data for the period 1948 to 1960, he fits alternative regressions including one, two, or three output (and relative price) terms, and selects the best-fitting equation satisfying *a priori* constraints on the coefficients. For the telephone industry, the best-fitting equation contains $\ln K_{t-1}$, t , $\ln Q_t$, $\ln Q_{t-1}$, and $\ln (w/c)_{t-1}$ and a constant term.

The values of the long-run elasticity of capital stock with respect to output implied by his regressions range from 1.0764 to 1.4319—the value for the best-fitting equation being 1.2289. He accepts the hypothesis of unitary output elasticity and estimates another regression function constraining the long-run elasticity to unity. The coefficients of time and $\ln (w/c)_{t-1}$ in the constrained function have correct signs and are significant at the 5-percent level. His analysis of the time profile of the investment response to a once-and-for-all increase of output indicates that the peak response occurs in the second year and that about 95 percent of the response is completed by the end of the third year.

□ **Jorgenson-Handel model.** The Jorgenson-Handel (J-H) econometric model of investment behavior⁷ has three components: (1) demand for capital, (2) replacement investment, and (3) the relationship between changes in demand for capital and actual investment.

The objective of the firm is to maximize its present value subject to the constraint that its output is given. Assuming that the production function has Cobb-Douglas form, that the factor markets are perfectly competitive, and that regulation is effective,⁸ their marginal productive condition for capital may be written:

$$\alpha(Q/K^+) = c/p,$$

where K^+ is the desired level of capital, Q is output, c is the rental price of capital, p is the price of output, and α is the elasticity of output with respect to capital stock.

They assume that replacement investment is proportional to capital stock,

$$I_t^R = \delta K_t,$$

where I_t^R is the level of replacement investment and δ is the rate of replacement.

The relationship between changes in demand for capital and

⁷ [5].

⁸ By effective regulation they mean that the actual price and output in a regulated industry are the same as those that would prevail under a competitive market structure.

actual investment expenditures is given by the following distributed lag function:

$$I_t - \delta K_t = \mu(S)[K_t^+ - K_{t-1}^+],$$

where $\mu(S)$ is a power series in the lag operator. Assuming that the sequence of coefficients $\{\mu_r\}$ of the distributed lag function has a rational generating function, the distributed lag function may be written as follows:

$$\omega(S)[I_t - \delta K_t] = \gamma(S)\alpha[(P_t Q_t/c_t) - (P_{t-1} Q_{t-1}/c_{t-1})],$$

where $\omega(S)$ and $\gamma(S)$ are polynomials in the lag operators.

They apply quarterly data from the OBE-SEC Investment Survey⁹ for the period 1949 to 1960. They use gross business income, the value of all capital input, as a measure of output. Their measure of rental price of capital, c , takes into account some aspects of the U. S. corporate tax structure.

They limit the order of the polynomial $\omega(S)$ to two, but allow the order of the polynomial $\gamma(S)$ to range up to 13 with about four non-zero coefficients. The best-fitting regression is chosen by minimizing the standard error of the regression, subject to the condition that the leading coefficient of the $\gamma(S)$ polynomial be nonnegative. For the communications industry, the best-fitting equation contains the following $\gamma(S)$ polynomial:

$$\gamma(S) = \gamma_7 S^7 + \gamma_8 S^8 + \gamma_9 S^9 + \gamma_{10} S^{10}.$$

The implied elasticity of output with respect to capital input is 0.24748. Their analysis of the time pattern of net investment that results from a change in the desired level of capital indicates that (1) no response in investment expenditures occurs in the first seven quarters, (2) the peak response occurs in the tenth quarter, and (3) the average lag is 14.03 quarters.

For purposes of comparison, the main features of these models are summarized in Table 1. These models have three components:

TABLE 1

COMPARISON OF ALTERNATIVE ECONOMETRIC MODELS OF INVESTMENT BEHAVIOR
IN TELEPHONE INDUSTRY

AUTHOR(S)	EXPRESSION FOR DESIRED CAPITAL	TIME STRUCTURE OF INVESTMENT PROCESS	REPLACEMENT INVESTMENT	TIME UNIT	TIME PERIOD
CLARK	$K_t^+ = \gamma \beta Q_t^e$	FIXED LAG OF ONE YEAR	PROPORTIONAL TO NET CAPITAL STOCK	YEAR	1912 TO 1948
HICKMAN	$K_t^+ = A (Q_t^e)^a \{(w/c)_t^e\}^b e^{ct}$	GEOMETRIC LAG DISTRIBUTION	PROPORTIONAL TO NET CAPITAL STOCK	YEAR	1949 TO 1960
JORGENSEN AND HANDEL	$K_t^+ = \alpha (p_t Q_t/c_t)$	RATIONAL DISTRIBUTED LAG FUNCTION	PROPORTIONAL TO NET CAPITAL STOCK	QUARTER	1949 TO 1960
SANKAR	$K_t^+ = A (p/c)_t^a Q_t^e [\sigma + (1-\sigma)/v]$	RATIONAL DISTRIBUTED LAG FUNCTION	PROPORTIONAL TO NET CAPITAL STOCK	YEAR	1949 TO 1968
K_t^+ = DESIRED CAPITAL Q_t^e = EXPECTED OUTPUT $(w/c)^e$ = EXPECTED WAGE-RENTAL RATIO			p = PRICE OF OUTPUT Q = OUTPUT c = RENTAL PRICE OF CAPITAL		

⁹ The OBE-SEC Survey is currently published in the March, June, September, and December issues of *Survey of Current Business* [10].

(1) determination of desired capital, (2) characterization of the time structure of the investment process, and (3) replacement investment. These models differ with respect to components (1) and (2), but they all assume that replacement investment is proportional to net capital stock.

Differences in the expressions for desired capital are due to differences in assumptions about the production function and the firm behavior. Clark assumes no substitution possibilities among inputs in the production process. Hickman and Jorgenson and Handel assume substitution possibilities in the production process. Hickman does not specify the production function. His expression for desired capital is of Cobb-Douglas form with different coefficients for expected output and expected relative factor price. Jorgenson and Handel assume a Cobb-Douglas production function which implies unitary elasticity of substitution between capital and labor. With respect to returns to scale, Clark and Jorgenson and Handel assume constant returns to scale. Hickman finds support for a constant returns to scale hypothesis and imposes this restriction on his final regression.

Hickman and Jorgenson and Handel assume that the objective of the firm is to minimize cost for given output. Jorgenson and Handel assume that current output is given—determined uniquely from the demand curve for the price fixed by the regulator. Hickman assumes that expected output and expected relative factor price are given. Expected price (relative factor price) is assumed to be a multiplicative function of current and two-lagged outputs (relative factor prices). Clark has no optimization problem, but he assumes that the firm’s investment decisions are governed by the firm’s expectation of future demand. In his model expected output in period t is a function of output in period $t - 1$ and output in period $t - 2$.

With respect to the time structure of the investment process, Clark assumes a fixed lag of one year, Hickman employs a geometric lag distribution, and Jorgenson and Handel employ a rational distributed lag function. The J-H function is more general in the sense that it includes the other two as special cases.

In his previous paper¹⁰ the author modified the J-H model by relaxing three of their assumptions: constant returns to scale, unitary elasticity of substitution, and equality between marginal cost and price. In his application to the U. S. electric utility industry he found that these assumptions are not valid for that industry. It is also desirable to test these assumptions with data relating to the U. S. telephone industry. In Section 3 he summarizes the model and compares it with the models surveyed in this section.

■ The production function for a telephone utility is assumed to be:

$$Q = A[dK^{-\rho} + (1 - d)L^{-\rho}]^{-1/\rho}, \tag{1}$$

where Q is output, K is capital, L is labor, A is the (neutral) efficiency parameter, d is the distribution parameter, ν is the returns to scale parameter, and ρ is the substitution parameter. The elasticity of substitution between capital and labor, σ , is $1/(1 + \rho)$. The restrictions

3. The model

on the parameters are:

$$A > 0, \quad 0 < d < 1, \quad \text{and} \quad v > 0, \quad \rho > -1. \quad (2)$$

As with Jorgenson and Handel,¹¹ the objective of the utility is assumed to minimize cost for a given level of output. The equation for total cost is

$$C = wL + cK, \quad (3)$$

where w is the wage rate and c is the rental price of capital. The marginal condition with respect to capital stock is¹²

$$c = v\lambda A^{-\rho/v} d Q^{1+\rho/v} K^{-1-\rho}, \quad (4)$$

where λ is the long-run marginal cost.

The regulatory agency is assumed to set the price of output equal to long-run average cost (including a fair return on capital). For the CES function long-run average cost is equal to the product of returns to scale (v) and long-run marginal cost (λ). Hence, equation (4) can be written as

$$c = p d A^{-\rho/v} Q^{1+\rho/v} K^{-1-\rho}, \quad (5)$$

where p is the price of output. The expression for desired level of stock is

$$\begin{aligned} K^+ &= (d A^{-\rho/v})^{1/(1+\rho)} Q^{(v+\rho)/v(1+\rho)} (p/c)^{1/(1+\rho)} \\ &= \text{const } (p/c)^\sigma Q^{[\sigma+(1-\sigma)/v]}. \end{aligned} \quad (6)$$

It may be noted that for the case of $\sigma = 1$, equation (6) is of the same form as the one derived by Jorgenson and Handel.

As with Eisner and Nadiri¹³ the relationship between changes in desired level of capital stock and actual changes in capital stock is assumed to be of the following form:

$$\ln(K_{t+1}/K_t) = \mu(S) \ln(K_t^+/K_{t-1}^+), \quad (7)$$

where $\mu(S)$ is a polynomial in the lag operator. Under the assumption that the sequence of coefficients $\{\mu_r\}$ of the distributed lag function has a rational generating function, equation (7) may be written:

$$\Delta \ln K_t = \frac{\gamma(S)}{\omega(S)} \Delta \ln K_t^+, \quad (8)$$

where $\gamma(S)$ and $\omega(S)$ are polynomials in the lag operators. Since we are employing annual data in our work, we think it is desirable to restrict the order of the $\omega(S)$ polynomial to two and the order of the $\gamma(S)$ polynomial to three.

For $\gamma(S)$ of order 3 and $\omega(S)$ of order 2, after equation (6) has been substituted into equation (8) and an error term ϵ_t has been

¹¹ The objective of the firm in the J-H model is to maximize its present value. However, under their assumptions of given output and no adjustment costs, their equilibrium conditions are similar to the conditions in a cost-minimization model.

¹² For derivation, see Sankar [8], p. 650.

¹³ In [2].

added, the distributed lag function may be written:¹⁴

$$\Delta \ln K_t = \sum_{s=0}^2 [\gamma_{ps} \ln (p/c)_{t-s} + \gamma_{qs} \Delta \ln Q_{t-s}] - \sum_{j=1}^2 \omega_j \Delta \ln K_{t-j} + \epsilon_t. \quad (9)$$

It is assumed that the error term is distributed independently and identically over time and distributed independently of all values of changes in desired capital stock. If the distributed lag function is also stable, the ordinary least-squares estimator of the unknown parameters is best, asymptotically normal.¹⁵

Certain restrictions on the coefficients of equation (9) have to be imposed so that the sequence of coefficients $\{\mu_r\}$ of the distributed lag function will be nonnegative.¹⁶ The restrictions are: (1) the leading coefficients of the $\gamma(S)$ polynomial be nonnegative, and (2) $4\omega_2 \leq \omega_1^2$.

It has been noted that Sankar's model generalizes the J-H model.¹⁷ It may appear that the models of Clark and Hickman are more general, at least in some respects, than this model, because investment decisions in their model are based on expected output. Although it is more realistic to assume that the entrepreneur makes investment decisions on the basis of expected output rather than on current output as in this model, it must be noted that it is very difficult in empirical work to separate a process which generates expectations from a process which generates an adjustment lag. For example, Hickman's regression model,

$$\ln (K_t/K_{t-1}) = \sum_{s=0}^3 a_s \ln Q_{t-s} + \sum_{i=0}^3 b_i \ln (w/c)_{t-i} + \beta \ln K_{t-1} + \text{other terms},$$

could be interpreted as either an expectation model (expected output—relative factor price—as a function of current and lagged outputs—relative factor prices) with a geometric lag distribution or a rational distributed lag function with $\gamma(S)$ of order 3 and $\omega(S)$ of order 1.

In Section 5 the reduced equation (9) is estimated with data relating to the U. S. telephone industry for the period 1949 to 1968. If the null hypothesis that the long-run elasticities of capital stock with respect to relative price and output are equal to unity is accepted, then the Cobb-Douglas specification of Jorgenson and Handel is valid.

■ The model developed in Section 3 is applied to annual data relating to the U. S. telephone industry. Regressions based on the lag

4. Data

¹⁴ In what follows the expression γ_{ps} stands for $\gamma_{(p/c)s}$. Also note that in equation (9) $\ln (p/c)_{t-2}$ corresponds to polynomial of order 3.

¹⁵ See [6], p. 181.

¹⁶ For details, see Jorgenson [4], pp. 146–147.

¹⁷ Although the author allows for the possibility that the coefficients of relative price and output in equation (6) are different, he imposes the *same* lag distribution for both variables. With only 20 observations, it is very difficult to get meaningful estimates if he introduces separate lag functions for the two variables.

specification in equation (9) have been fitted to data for the period 1949 to 1968. The data for 1946 to 1968 were also used in measuring lagged variables. A description of the sources of data and of the measurement of variables used in regression analysis is given below.

□ **Capital stock (K).** Under the assumption that replacement investment is proportional to net capital stock, it is possible to construct capital stock series with (1) an estimate of the initial capital stock, (2) the depreciation rates, and (3) the series on gross investment.

The estimate of the initial capital stock is based on Ulmer's work.¹⁸ His estimate of the value of plant and equipment (as of January 1, 1946) in current dollars is \$3063.8 millions. This amount was deflated by using the investment goods price index, and this resulted in an estimate of \$3784 millions.

Depreciation rates for the consolidated companies in the Bell System are given in Hickman and in Moody's *Public Utility Manuals*.¹⁹ These depreciation rates are on the straight-line basis, which is the method prescribed by the Uniform System of Accounts. Hickman's rates were used for the years 1946 to 1949, and for later years, the rates from Moody's *Public Utility Manuals*. It is assumed that these rates are valid for the industry.

Data relating to construction expenditures in U. S. telephone industry (excluding Alaska and Hawaii) and the composite telephone price index for plant and equipment were obtained from the American Telephone and Telegraph Company.

Finally, the capital stock series (as of January 1) in 1954 dollars was computed with the following formula

$$K_{t+1} = I_t + (1 - \delta_t)K_t,$$

where I_t is construction expenditure in year t in 1954 dollars and δ_t is the depreciation rate for year t .

□ **Price index for investment goods (g).** The composite telephone price index for plant and equipment (1954 = 100), provided by the American Telephone and Telegraph Company, was used.

□ **Outprice price (p).** The price deflator for telephone communication provided by the Office of Business Economics, U. S. Department of Commerce, was used.

□ **Output (Q).** An FCC publication contains data relating to operating revenues.²⁰ By dividing operating revenue of each year by the output price of that year, a measure of telephone output is obtained.

□ **Interest rate (r).** Moody's weighted average of yields on newly issued domestic bonds and preferred stocks was used.

□ **Rental price of capital (c).** $c_t = (r_t + \delta_t)g_t$.

Data relating to capital stock, output, output price index, interest rate, depreciation rate and investment goods price index are given in Table 2.

¹⁸ See [9].

¹⁹ [7].

²⁰ See [11], Table 9, p. 20.

TABLE 2

TELEPHONE INDUSTRY: VALUE OF CAPITAL STOCK, OUTPUT, OUTPUT PRICE INDEX, INTEREST RATE, DEPRECIATION RATE, AND INVESTMENT GOODS PRICE INDEX

YEAR	(1)	(2)	(3)	(4)	(5)	(6)
	VALUE OF CAPITAL STOCK (MILLIONS OF 1954 DOLLARS)	OUTPUT	OUTPUT PRICE INDEX (1954 = 100)	INTEREST RATE (PERCENT)	DEPRECIATION RATE (PERCENT)	INVESTMENT GOODS PRICE INDEX (1954 = 100)
1946	3,784	2,768	79.9	2.72	4.25	81.0
1947	4,524	2,903	81.2	2.77	4.26	93.7
1948	5,676	3,278	84.6	3.09	4.20	95.4
1949	7,064	3,496	87.4	3.11	4.14	95.5
1950	7,976	3,707	92.9	2.73	3.94	96.8
1951	8,638	4,036	94.6	3.38	3.88	103.3
1952	9,397	4,356	97.1	3.44	3.88	101.4
1953	10,363	4,618	100.2	3.75	3.84	98.9
1954	11,471	5,013	100.0	3.09	3.83	100.0
1955	12,557	5,609	99.2	3.30	3.87	101.2
1956	13,909	6,109	100.2	3.78	3.92	104.8
1957	15,808	6,533	101.7	4.62	4.58	107.5
1958	17,816	6,862	104.1	4.19	4.57	108.6
1959	19,336	7,384	105.6	5.13	4.64	109.8
1960	20,838	7,858	106.5	4.76	4.66	110.2
1961	22,691	8,343	106.7	4.63	4.70	109.5
1962	24,532	8,930	106.6	4.55	4.82	108.9
1963	26,592	9,542	106.5	4.29	4.88	110.3
1964	28,716	10,306	106.2	4.55	4.98	111.8
1965	31,087	11,303	104.1	4.61	5.10	112.0
1966	33,746	12,632	102.3	5.51	5.14	114.7
1967	36,462	13,464	103.0	5.77	5.21	118.9
1968	39,015	14,789	102.1	6.55	5.29	125.4
1969	41,596					

■ Within the lag specification in (9) it is possible to run several regression equations. The procedure in the selection of the final regression is essentially the same as the one used by Jorgenson and Stephenson.²¹ Sankar runs several regressions corresponding to different lag structures and chooses the one which minimizes the standard error of the regression subject to the constraint that the leading coefficient of the $\gamma(S)$ polynomial be nonnegative. If the best-fitting equation does not satisfy the additional condition, $\omega_2 \leq \omega_1^2/4$, then the constraint is imposed by taking the unconstrained estimator of ω_1 as an estimator of ω_1 and ω_2 as $\omega_1^2/4$.

Eight regression functions, corresponding to different lag specifications, satisfy the constraint that the leading coefficient of the $\gamma(S)$ polynomial be nonnegative. For these regression functions, information pertaining to the standard error and elasticities of capital with respect to relative price and output are given in Table 3. The relative price elasticities range from 0.02335 to 0.37046. All the output elasticities are above one, implying that the returns to scale parameter, ν , is less than one.

The best-fitting equation with a geometric distributed lag function ($\gamma_1, \gamma_2, \omega_1$) has a standard error of 0.01417. The regression func-

5. Results

²¹ In [6].

TABLE 3

STANDARD ERROR OF REGRESSION EQUATION (9), LONG-RUN ELASTICITIES OF CAPITAL STOCK WITH RESPECT TO RELATIVE PRICE AND OUTPUT FOR ALTERNATIVE FORMS OF THE POLYNOMIALS $\gamma(S)$ AND $\omega(S)$

(1)	(2)	(3)	(4)	(5)
$\gamma(S)$	$\omega(S)$	STANDARD ERROR (CORRECTED FOR DEGREES OF FREEDOM)	ELASTICITY OF CAPITAL WITH RESPECT TO RELATIVE PRICE	ELASTICITY OF CAPITAL WITH RESPECT TO OUTPUT
0	1, 2	0.01454	0.02335	1.10434
1	1	0.01622	0.09934	1.11325
1	1, 2	0.01637	0.12375	1.12399
2	1	0.01663	0.33179	1.06223
2	1, 2	0.01470	0.31863	1.15498
0, 1	1, 2	0.01481	0.07963	1.14467
1, 2	1	0.01417	0.24247	1.16698
1, 2	1, 2	0.01196	0.37046	1.24847

tion is²²

$$\begin{aligned} \Delta \ln K_t = & 0.06934 \Delta \ln (p/c)_{t-1} + 0.08972 \Delta \ln (p/c)_{t-2} \\ & (0.04264) \quad (0.04297) \\ & + 0.50780 \Delta \ln Q_{t-1} + 0.25773 \Delta \ln Q_{t-2} + 0.34401 \Delta \ln K_{t-1} \\ & (0.18324) \quad (0.11870) \quad (0.12290) \end{aligned}$$

The implied estimates of the elasticities of capital with respect to relative price and output are 0.24247 and 1.16698 respectively.

The best-fitting equation subject to the constraint that the leading coefficient of the $\gamma(S)$ polynomial be nonnegative is

$$\begin{aligned} \Delta \ln K_t = & 0.10384 \Delta \ln (p/c)_{t-1} + 0.10883 \Delta \ln (p/c)_{t-2} \\ & (0.03828) \quad (0.03699) \\ & + 0.33225 \Delta \ln Q_{t-1} + 0.38446 \Delta \ln Q_{t-2} + 0.76573 \Delta \ln K_{t-1} \\ & (0.16827) \quad (0.11102) \quad (0.18985) \\ & - 0.33981 \Delta \ln K_{t-2} \\ & (0.12806) \end{aligned}$$

$$s = 0.01196 \quad R^2 = 0.54584.$$

The coefficient of multiple correlation, R^2 , and the standard error of regression, s , are corrected for degrees of freedom. This regression function does not satisfy the condition, $\omega_2 \leq \omega_1^2/4$. Hence, the regression was reestimated with the unconstrained estimate of $\omega_1 = -0.76573$ and ω_2 as $\omega_1^2/4 = 0.14659$. The resulting regression function is

$$\begin{aligned} \Delta \ln K_t = & 0.76573 \Delta \ln K_{t-1} + 0.14659 \Delta \ln K_{t-2} \\ = & 0.06590 \Delta \ln (p/c)_{t-1} + 0.12879 \Delta \ln (p/c)_{t-2} \\ & (0.04081) \quad (0.03921) \\ & + 0.15543 \Delta \ln Q_{t-1} + 0.29484 \Delta \ln Q_{t-2} \\ & (0.11377) \quad (0.11614) \end{aligned}$$

Estimates of long-run elasticities of capital stock with respect to

relative price and output are:

$$\hat{E}_{(p/c)} = \hat{\sigma} = \frac{\hat{\gamma}_{p1} + \hat{\gamma}_{p2}}{1 + \hat{\omega}_1 + \hat{\omega}_2} = 0.51119$$
$$\hat{E}_Q = \left(\hat{\sigma} + \frac{1 - \hat{\sigma}}{\nu} \right) = \frac{\hat{\gamma}_{q1} + \hat{\gamma}_{q2}}{1 + \hat{\omega}_1 + \hat{\omega}_2} = 1.18225.$$

These point estimates indicate that the long-run elasticity of capital stock with respect to relative price is less than one and the long-run elasticity of capital stock with respect to output is more than one (and hence the implied point estimate of the scale parameter, ν , is less than one). However, it is not possible to reject either the hypothesis $\sigma = 1$ or $\nu = 1$ at the 5-percent level of significance. With respect to the scale parameter the result is similar to that of Hickman. Thus it appears that the Cobb-Douglas specification of Jorgenson and Handel is valid for the telephone industry.

The responses of net investment to relative price and output changes are given in Table 4 and in Figure 1. It can be seen from

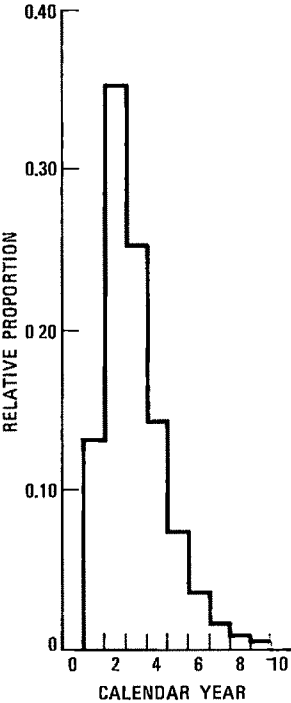
TABLE 4
DISTRIBUTED LAG COEFFICIENTS: RESPONSES OF INVESTMENT
TO RELATIVE PRICE AND OUTPUT CHANGES

$\Delta \ln K_t = 0.06590 \Delta \ln (p/c)_{t-1} + 0.12897 \Delta \ln (p/c)_{t-2} + 0.15543 \Delta \ln Q_{t-1}$ $+ 0.29484 \Delta \ln Q_{t-2} + 0.76573 \Delta \ln K_{t-1} - 0.14659 \Delta \ln K_{t-2} + \hat{\epsilon}_t$ $= \sum_{i=1}^{\infty} [\hat{\mu}_{pi} \Delta \ln (p/c)_{t-i} + \hat{\mu}_{qi} \Delta \ln Q_{t-i}] + \hat{\epsilon}_t$						
(1)	(2)	(3)	(4)	(5)	(6)	(7)
i	$\hat{\mu}_{pi}$	$\sum_{k=1}^i \hat{\mu}_{pk}$	(3)/ $E_{(p/c)}$	$\hat{\mu}_{qi}$	$\sum_{k=1}^i \hat{\mu}_{qk}$	(6) / E_Q
1	0.06590	0.06590	0.12891	0.15543	0.15543	0.13147
2	0.17925	0.24515	0.47957	0.41386	0.56929	0.48153
3	0.12760	0.37275	0.72918	0.29413	0.86342	0.73032
4	0.07143	0.44418	0.86891	0.16455	1.02797	0.86950
5	0.03600	0.48018	0.93934	0.08288	1.11085	0.93961
6	0.01710	0.49728	0.97279	0.03934	1.15019	0.97288
7	0.00781	0.50509	0.98807	0.01797	1.16816	0.98808
8	0.00347	0.50856	0.99486	0.00799	1.17615	0.99484
9	0.00152	0.51008	0.99783	0.00349	1.17964	0.99779
10	0.00065	0.51073	0.99910	0.00150	1.18114	0.99906
11 TO ∞	0.00046			0.00111		1.00000
SUM	0.51119		1.00000	1.18225		
MODAL LAG	2.00000			2.00000		
MEAN LAG	2.90143			2.89544		

columns (2) and (3) that the peak responses occur in the second year. Hickman also finds the peak response for output changes in the second year, but in his case the relative weight for the second year is much larger. The estimate of the average lag (2.9 years) is closer to that of Hickman but smaller than that of Jorgenson and Handel (3.5 years).

In an earlier paper,²³ results for the electric utility industry were

FIGURE 1
THE TIME FORM OF LAGGED
RESPONSE OF NET INVESTMENT
TO CHANGE IN RELATIVE PRICE
OR OUTPUT



²³ [8].

reported. Regression results based on c_1 as a measure of rental price of capital for the electric utility industry can be compared with the results for the telephone industry.²⁴ For the electric utility it was found that the elasticities of capital with respect to relative price and output are 0.18317 and 0.64573, respectively. Also, the implied estimate of the scale parameter, v , was shown to be significantly above unity, and the hypothesis of zero elasticity of substitution between capital and labor could not be rejected at any reasonable level of significance. Now, for the telephone industry the Cobb-Douglas specification with a constant returns to scale constraint is found to be valid. The time pattern of net investment that results from a change in relative price or output also differs for these two industries. In comparing Table 3 of the previous paper²⁵ with Table 4 of this paper the following differences may be noted. (1) For the electric utility industry the first period responses are zero, whereas for the telephone industry the first period responses are about 13 percent of the totals. (2) For the telephone industry only about 6 percent of the total responses remain to be completed after the end of the fifth year, whereas the corresponding percentages for the electric utility industry are above 18 percent. (3) The average lag for the electric utility industry exceeds the average for the telephone industry by more than a year.

In his earlier paper the author experimented with three different measures of the rental price of capital. His measures, c_2 and c_3 , incorporated the U. S. tax structure. Unfortunately, he could not carry out similar experiments for the telephone industry as a result of lack of necessary information. With respect to depreciation policy, the Bell System used the straight-line method during the sample period, so the author has assumed the same practice for other firms as well. All firms in the industry took advantage of investment tax credits, but he could not find information about the actual rates.²⁶ Hence both accelerated depreciation and investment tax credits were ignored, and the following measure of the rental price variable was constructed,

$$c_{3t} = g_t(r_t + \delta_t)(1 - u_t Z_t)/(1 - u_t),$$

where Z is the present-value of depreciation deduction (straight-line basis) on one dollar's investment, and u is the marginal tax rate for corporate income. Using this measure, he estimated equation (9) corresponding to different lag specifications. The results are similar to the ones based on the measure c , but the least-squares fits are generally poor. The best-fitting equation, subject to the constraint that the leading coefficient of the $\gamma(S)$ polynomial be nonnegative, corresponds to the same lag specification as before, but the standard error of the regression is 0.01548, with R^2 (corrected for degrees of freedom) only 0.24006. Hence, he does not present additional results for these regressions.

²⁴ The only difference between c_1 and c is that c_1 is based on a constant depreciation rate, whereas c is based on varying annual depreciation rates.

²⁵ [8].

²⁶ The allowed rate is generally about 3 percent of new depreciable assets.

■ In this paper the author has applied an earlier developed model²⁷ to the U. S. telephone industry for the period 1949 to 1968. He finds that the long-run elasticities of capital with respect to relative price and output are 0.51119 and 1.18225, respectively. These estimates imply that the elasticity of substitution between capital and labor is 0.51119 and the returns to scale parameter is less than one. However, neither the hypothesis that the elasticity of substitution between capital and labor is equal to one nor the hypothesis that the returns to scale parameter is equal to one can be rejected at any reasonable level of significance. In analyzing the time pattern of investment that results from a change in the level of relative price or output, an estimate of 2.9 years is found for the average lag.

In comparing these results with those for the electric utility industry, important differences may be noted. (1) For the electric utility industry the estimate of the scale parameter is significantly above one and the estimate of the elasticity of substitution is close to zero, and hence the Cobb-Douglas specification is inappropriate, whereas for the telephone industry the Cobb-Douglas specification appears to be valid. (2) The average lag for the electric utility industry exceeds that of the telephone industry by more than a year. (3) The first-period response is zero for the electric utility industry, whereas the relative first-period response for the telephone industry is 13 percent. These differences—both in technology and in time structure of net investment—suggest that it is not meaningful to treat public utilities as one group for the purpose of analyzing investment behavior.

In both papers the author has applied his model at the industry level. This enabled him to tap new data sources and to construct suitable measures of the variables. He found some difficulties in developing an appropriate measure of the rental price of capital variable to incorporate the effects of tax policies. On the estimation side, his main concern is that what he has is only large-sample results. The large-sample properties are cold comforts to an econometrician working with only 20 observations and 6 explanatory (including lagged dependent) variables.

²⁷ In [8].

References

1. CLARK, P. G. "The Telephone Industry: A Study in Private Investment," in W. Leontieff, ed., *Studies in the Structure of the American Economy*, New York: Oxford Univ. Press, 1953.
2. EISNER, R. AND NADIRI, M. "Investment Behavior and Neo-Classical Theory," *The Review of Economics and Statistics*, Vol. 50, No. 3 (August 1968), pp. 369-382.
3. HICKMAN, B. G. *Investment Demand and U. S. Economic Growth*. Washington, D. C. The Brookings Institution, 1965.
4. JORGENSEN, D. W. "Rational Distributed Lag Functions," *Econometrica*, Vol. 34, No. 1 (January 1966), pp. 213-264.
5. ——— AND HANDEL, S. S. "Investment Behavior in U. S. Regulated Industries," *The Bell Journal of Economics and Management Science*, Vol. 2, No. 1 (Spring 1971), pp. 213-264.
6. JORGENSEN, D. W. AND STEPHENSON, G. A. "Investment Behavior in U. S. Manufacturing, 1947-60," *Econometrica*, Vol. 35, No. 2 (April 1967), pp. 169-220.

7. MOODY, JOHN. *Moody's Public Utilities Manuals*. New York: John Moody, 1971.
8. SANKAR, U. "Investment Behavior in the U. S. Electric Utility Industry, 1949-1968." *The Bell Journal of Economics and Management Science*, Vol. 3, No. 2 (Autumn 1972), pp. 645-664.
9. ULMER, M. J. *Capital in Transportation, Communication, and Public Utilities*. Princeton: National Bureau of Economics Research, 1960.
10. U. S. DEPARTMENT OF COMMERCE, OFFICE OF BUSINESS ECONOMICS. *Survey of Current Business*, various monthly issues.
11. U. S. FEDERAL COMMUNICATIONS COMMISSION. *Statistics of Communications Common Carriers*, Year Ended December 31, 1969.

Ralph Turvey's Economic Analysis and Public Enterprises

Reviewed by

William G. Shepherd

Professor of Economics
University of Michigan

■ This remarkable, slim volume¹ is really two extra-slim volumes in one. The first is a prismatically lucid treatment of economic criteria for public enterprises (and indeed for private firms under regulation). The second applies these criteria to a variety of actual enterprises in Britain, including gas, coal, and urban public transport.

The volume arises from Turvey's experience as (1) chief economist for the British Electricity Council in the mid-1960s, and then (2) member and deputy chairman of the U.K. Prices and Incomes Board (PIB) from 1967 to 1971. In the former role he recast the economic guidance of the British electricity system² and touched all major theoretical aspects of utility economics—public or private. On the PIB—which screened price increases in the British economy from 1966 to 1971, and rejected some!—he further developed the criteria and applied them to a wide range of public and private firms, in a variety of monopoly and quasicompetitive conditions.

With his proven talent for making the most of a topic in the least space, Turvey has created required reading for everyone in the field, academic and practical alike. This also includes those who manage, regulate, or study regulated utilities in the United States. And the book's brevity and clarity make the subject more accessible than I had thought possible.

The first little book (66 pages in 7 chapters) presents and explains the criteria for optimizing prices and investment. From the simple premise, that the aim is to maximize *Social Benefit* minus *Social Cost*, Turvey develops increasingly sophisticated criteria for the basic decisions that must be made. Simple rules of Financial Performance are followed by Second Best (Chapter 3). Flow charts, mathematics, and practical examples make the strange pricing lessons of second best (set price above, at, or below marginal cost, depending on several conditions) remarkably clear and intuitively satisfying. The merits of Simple versus Complex Price Structures follow. Here, as in the other "theoretical" chapters, Turvey uses practical examples adroitly and in abundance. Postal, water, bus, telephone—the whole range of familiar "utility" services—are called in to show what the concepts are about.

Next comes Service Quality, particularly the choices to be made when service quality has several dimensions. Then follows an ex-

¹ R. Turvey, *Economic Analysis and Public Enterprises* (London: George Allen & Unwin, 1971).

² See his *Optimal Pricing and Electricity Supply* (Cambridge: M.I.T. Press, 1968), which complements his present book.

tended treatment of those crucial Marginal Costs (Chapter 6). The analysis is for an ongoing enterprise with growing demand, in which new capacity is cheaper to run than older capacity. Marginal costs are time-related, and although they are complex to define, they may be amenable to rough approximations in practice. This leads into Optimal Pricing over Time (Chapter 7), in which Turvey develops a "short-term pricing rule" and a "long-run investment rule" as parts of the strategy for maximizing Benefit-less-Cost.

The case studies in the second half (67 pages in 6 chapters) are drawn from several PIB reports in which Turvey played the main role. The details are British. Yet the lessons are not provincial, and indeed American readers will find that the facts are fresh and the policy analysis is lively and full of implications for U. S. practices.

In the main, Turvey works within the context that public enterprises' primary task is to operate on sound commercial lines. At most points, price is taken to measure social benefit. This fits the recent trends in British thought and practice. But Turvey does not carry this to extremes, and indeed the analysis can be adapted to include most of the important social elements that are likely to arise in public enterprises — except, as Turvey notes, equity.

Practical experts will credit the book's balance, brevity, and clarity to the pressures for real solutions under which Turvey worked at the Electricity Council and at the PIB. Academic tomes on public enterprise, by contrast, are often too long and soft, or tendentious. Yet without Turvey's analytical skill, the surroundings probably would not have helped. Perhaps only temporary exposure works; long practical experience often seems only to lower the plane of analysis.

Despite its title, the book's content is almost entirely applicable to optimizing in private firms under regulation. The institutional setting differs, and regulation may also lay on additional, or even contrary, inducements affecting utility behavior. But the core of normative analysis which Turvey presents has generality. And on marginal-cost pricing, in particular, Turvey's analysis could and should advance American discussions appreciably.

Apart from its specific instructional value in clarifying criteria, the book carries the basic message that theory can be both clear in content and useful in hard-pressed practical situations. With skill and good sense, one can indeed rough out the essential optimizing factors and then deal with them. This is a powerful lesson. But it needs to be kept in proportion. The analytical tools are still few and simple, and on many key points the issues are still open. That is one reason that Turvey's real but finite contribution seems so unusually valuable. Turvey's work is not an open invitation for vast or mechanical solutions. On the contrary, his best results are in direct studies of limited problems, done with a sure touch.

The very generality and elegance of Turvey's analysis may give some readers pause. Can such a comprehensive and flexible approach be made operational? Can those really tricky public interest problems — of redundant coal miners, of a "Buy British" policy for aircraft and electric generators, etc. — be inserted and solved as easily as Turvey indicates?

The answer, I think, is essentially yes, for most public enterprises. Turvey's analysis starts on the most general plane, by assuming away such public-interest complications. Precisely because of this clean-

ness, the analysis is mostly transferable to regulated private firms. And the approach is operational to a first approximation for most British public enterprises. Only the few remaining "sick-industry" cases (coal, railroads?) have social impacts which probably require explicit inclusion in the optimizing process. For the rest, Turvey provides the basic road map, which needs only ad hoc adjustments in certain areas. Maximizing social benefit comes remarkably close to finding and approximating Pareto conditions. Therefore, Turvey's approach puts first things first, on the whole. And toward the end he also gives cases where social aspects *are* included.

And finally, Turvey's analysis abstracts from those hard problems of competition and structure which presently suffuse many regulated utility sectors in the U. S. (railroads, airlines, telecommunications, bulk electricity, cable TV, etc.). They also affect many public enterprises, too, in Britain and elsewhere. *Given* structure, Turvey clarifies the criteria. But what the optimum structure is, and how it may evolve, Turvey sets aside as a separate problem. Perhaps he will come to illumine it, in time. One hopes so. Meanwhile, he has given us an excellent introduction to the uses of economic analysis for optimizing in public enterprises, and in other settings too.

George C. Eads' The Local Service Air Line Experiment

Reviewed by

Philip Verleger

Data Resources, Incorporated
Lexington, Massachusetts

■ *The Local Service Air Line Experiment* by George C. Eads¹ reviews the costs and benefits of the program begun in 1945 by the United States Congress to provide scheduled air service to smaller cities. The author attempts to provide a detailed analytical discussion of local service operation as well as a lengthy analysis of the quality of management both by the U. S. Civil Aeronautics Board (CAB) and by local service operators. Unfortunately, the book does not fully achieve the author's goals, since the presentation is frequently rushed and consequently incomplete, especially in the analytical sections. However, in his concise analysis of the institutional history of the local service experiment, Eads' arguments are both complete and convincing.

The book is organized into three sections. The first, Chapters 2 and 3, analyzes the cost and demand characteristics of local airline service. The second section, Chapters 4, 5, and 6, presents a review of the historical pattern of CAB regulation of local service carriers. The final section, Chapters 7 and 8, analyzes the quality and costs to the public of the government's attempt to provide efficient short-haul service to communities which had not previously received it, and presents the reader with a short list of policy alternatives which Eads feels the government should now consider.

In the first section, Eads dives into the question of local service demand and cost functions. Chapter 2 presents a brief examination of demand covering only the demand for "short-haul" air service (not local service, although one is evidently to presume that the two types are roughly synonymous). Chapter 3 contains a complete analysis of local service costs. The author presents statistical equations for eight categories of expenses to emphasize the basic point—that by selecting larger rather than smaller aircraft, the airlines placed themselves in a situation where they could not afford to service the smaller cities they were chartered to serve.

Section two contains the main thrust of the author's argument. In Chapter 5, he describes the "local service experiment" along with the historical background and the abrupt manner in which the experiment ended when local service carriers' certificates, which had been temporary, were made permanent. At the same time the subsidy became guaranteed. Eads dwells at length on the CAB-created problems of determining the amount of subsidy and the initial restrictions on local service routes, which required every flight to stop

¹ G. C. Eads, *The Local Service Air Line Experiment* (Washington, D. C.: The Brookings Institution, 1972).

at every city. The later relaxation of this restriction enabled the local carriers to compete better with the trunk carriers, and in turn accelerated their choice of larger over smaller aircraft. The choice was further dictated by the Board's transfer of routes from trucks to locals. Eads goes on to show that, by allowing these equipment acquisitions, the CAB caused the local service experiment to fail because the larger planes could not provide the small cities with sufficient flight frequencies to tap fully potential demand.

The acquisition of trunk line piston equipment as well as jets preceded graduation by the local carriers into regional trunk line status. This transition is documented in Chapter 7, which describes the CAB's final efforts to make the locals self-supporting. These efforts entailed granting bypass authority and new routes to the locals so that they came into direct competition with the trunks. In addition, several mergers between carriers were allowed. The net effect of these and earlier actions of the Board was to downgrade further the service received by smaller cities. This degradation in service is discussed in Chapter 7 where Eads summarizes his findings by stating:²

The Board has never provided the local carriers with a sufficient incentive to use the kind of aircraft that would allow them to offer adequate service to their smaller stations. Indeed, with the limited exceptions noted above in Chapter 5, it has, through its route and subsidy policies, allowed or even actively encouraged the local service carriers to acquire aircraft that are suitable only for dense, longer-haul routes. Its removal of the intermediate stop restrictions, though necessary, was done in a way that encouraged competition between the local service carriers and the trucklines, rather than allowing the local service carriers the flexibility needed to serve their smaller points adequately. The Board's subsidy policy during the pre-class rate period actively encouraged the use of excessively large aircraft, and the class rate subsidy system, even when it has been "neutral" as among types of aircraft, has been administered in such a way as to encourage service cutbacks at the smaller stations.

Eads concludes by offering four prescriptions for the industry: (1) eliminate the subsidy; (2) compensate the carriers for serving small isolated cities; (3) encourage the locals to contract with small air taxis to provide inexpensive, frequent service; and/or (4) have the government contract directly with the air taxis.

Through his analysis, Eads attempts to make two points. The first is that the "local service experiment" failed due to the administrative errors of the CAB, particularly in its handling of the subsidy. Here he succeeds admirably. Eads makes his case by presenting a detailed investigation of the demands and costs of local airline service and by evaluating the effects of CAB decisions on the local carriers. Through the analysis, he first indicates that demand is greater for frequent flights which, due to the market size of many stations, can be satisfied at a profit only if aircraft with less than twenty seats were used. However, the local carriers bought much larger equipment, and thus generally provided less frequent service to smaller stations, usually at a loss. Eads indicates that this incorrect equipment choice could have been corrected by the CAB through its allocation of subsidy money. In his view, the Board's unwillingness to exercise sufficient control in subsidy allocation thus becomes one of the major causes of the failure of the local service experiment because,

² P. 183.

by not discouraging the purchase of larger than optimally sized aircraft, any possibility of profitable local airline service to smaller cities was ended. He goes on to argue that the local carriers continued to provide unprofitable service only so long as subsidies in amounts to cover the losses were provided. And, as the Board sought in the late 1960s to reduce subsidies, Eads points out that these services were inevitably canceled.³

Eads' second, and more important, point is that the experiment could have succeeded had the Board used its authority over the distribution of subsidy monies as well as its charter granting authority to persuade the local service airlines to provide frequent service with suitable equipment (i.e., small "air taxi" type of aircraft). On this point, Eads is not convincing because he does not provide a satisfactory analysis of the determinants of *local service demand*, choosing to rely on the results of others in spite of the fact that the works he cites were done primarily for traffic between large city-pairs rather than the smaller cities which are the author's concern. While this is not the proper forum in which to evaluate the demand studies the author has selected, it is important to note that throughout the book the author asks the reader to accept the premise that many local service cities would have generated more traffic had they been given frequent service by small aircraft. Since neither of the two primary studies cited by the author⁴ examined the effect on demand of the aircraft substitution prescribed by the author, one finds it difficult to accept Eads' assertion. The author's position is particularly weakened by the fact that many of the local service cities with infrequent service in the 1960s were, according to Eads, within a two hours' drive of an airport serving a larger city. In short, there would seem to be an obligation to demonstrate that there was, in fact, sufficient local service demand to warrant different CAB policies on subsidies and route awards.

Thus, while the book presents a convincing and detailed analysis of another regulatory failure, the author falls short of selling his policy prescriptions due to the inadequate analysis of demand. The book is, however, worthwhile reading for anyone interested in the impact and effect of regulation, as it provides a comprehensive and educational analysis of the historical (administrative) causes for this regulatory failure. One hopes, though, that later editions will provide the necessary analysis of local service demand to illustrate, beyond question, that the CAB misused its regulatory authority, at the expense of both the local carriers and the consumers. This task has been left for the reader.

³ This theme is emphasized throughout the book, but especially in Chapter 7.

⁴ Simpson [2] and Gronau [1].

References

1. GRONAU, R. *The Value of Time in Passenger Transportation: The Demand for Air Travel*. Occasional Paper 9, Columbia University Press, for National Bureau of Economic Research, 1970.
2. SIMPSON, R. W. AND NEUVE EGLISE, M. J. "A Method for Determination of Optimum Vehicle and Frequency of Service for a Short-Haul V/STOL Air Transport System." Report R-68-1, Massachusetts Institute of Technology, Flight Transportation Laboratory, May 1968.

The FTC staff's Economic Report on Conglomerate Merger Performance

Reviewed by

J. Fred Weston

Professor of Business Economics and Finance
University of California, Los Angeles

■ This second conglomerate merger Report¹ represents the completion of the second phase of a study begun by the staff of the Federal Trade Commission in 1968.² The first Report, issued in August 1969, represented a summary of more general evidence already in existence, and was not pinpointed to conglomerate mergers.³ The present Report is based on compilations of data developed in response to inquiries and derived from field investigations of a sample of nine conglomerate firms. Over fifty tables of mostly original data are developed to test a number of hypotheses about the effects of conglomerate merger activity during the 1960 to 1969 period. While this second study does not have the adversary tone of the first merger Report by the FTC staff, a relatively heavy emphasis on the potential harm from conglomerate merger remains. It is surprising that the additional data and materials on the operations of individual conglomerates, set forth in the seven volumes of the Celler Committee Hearings through May 1970, were not utilized.⁴

■ The Report provides valuable new information on market shares of acquisitions and resulting effects on market positions of the acquirers. The Report finds that 82 percent of the acquisitions by the conglomerates, from 1963 to 1969, represented a "toe-hold" acquisition (market share less than 5 percent).⁵ Indeed, 54 percent of the acquisitions represented five-digit product classes in which the market share in the year prior to acquisition was less than one percent. Further, the post-acquisition changes in market positions acquired by conglomerates were predominantly decreases.⁶ This pattern was not greatly influenced by the level of concentration of the market in which the position was taken. After the massive merger activity of the 1960s, the market share of the conglomerate firms in 1969 was less than 5 percent in 82.4 percent of the five-digit product classes in which they operated. The pattern is similar for market positions of less than one percent.⁷

1. Introduction

2. Data on market structure effects

¹ [12], An Economic Report to the Federal Trade Commission by the Bureau of Economics, H. Michael Mann, Director. Stanley E. Boyle of Virginia Polytechnic Institute and Philip W. Jaynes are authors of the study.

² [13].

³ *Ibid.*, p. xv.

⁴ See Celler [5, 6].

⁵ See [12], p. 121.

⁶ *Ibid.*, p. 123.

⁷ *Ibid.*, p. 117.

While the Report does not develop the implications of such data, it does acknowledge that the overall impact on competition due to conglomerate mergers was not appreciable. But even stronger implications may be ventured. In 98 percent of the acquisitions, a change in market position took place—either an increase or decrease. This high frequency of changes in market share positions does not support the mutual forbearance theory.⁸ A great emphasis of the first merger Report was on the increase in macroconcentration. This issue is not pursued in the second Report—analysis avoiding “statistical games” shows that a given set of the largest firms simply maintains its position in manufacturing industries over most time periods.⁹

3. Effects of conglomerates on business conduct

■ The Report assembled data on the extent to which the administration of individual functions in acquired companies was changed. Changes in a high percentage of cases were observed in (1) auditing, (2) insurance, and (3) financing. But in all other activities, little change of administrative function took place. For example, the responsibility for the advertising function was altered in only about 9 percent of the instances and predominantly the amount spent on advertising and promotion was unchanged. This finding contradicts the deep-pocket theory that the large conglomerate after acquisition will use financial power to dominate the new product areas by spending massive amounts on advertising.

Nor is the reciprocity argument supported by the data. Among the functions analyzed, the least number of administrative changes was observed in purchasing—about 4 percent of the acquisitions covered. The continued decentralization of the purchasing function and the emphasis on profit center responsibility of managers is inconsistent with the allegation that a strong motive for conglomeration was to facilitate reciprocity powers of the conglomerate firm.¹⁰ Thus the new evidence does not support the earlier apprehensions expressed toward conglomerate mergers.

4. Effects on efficiency of operations

■ If the conglomerate did not shift administrative functions, it is argued that it was not likely to contribute to increases in efficiency of the acquired entity. The Report further asserts that potentials for increasing efficiency are unlikely in view of Williamson's concept of multidivisional (*M*-form) method of operating multiunit activities.¹¹ But neither the *M*-form hypothesis of Williamson nor the interpretation of it assayed by the Report is inconsistent on theoretical or empirical grounds with contributions to efficiency by conglomerate mergers.¹² A number of companies in the Celler Committee Hearings described their emphasis on management, planning, and control processes for developing a communication system for improving the performance of operating entities.¹³

Tabulating whether the location of the performance of functions

⁸ See [13], pp. 458–471.

⁹ See Bock [4], pp. 17–18.

¹⁰ Cf. Merkhams [9].

¹¹ See Williamson [18].

¹² See Williamson [19].

¹³ See [6], pp. 91, 150, 160, and 380–381.

in a company was shifted misses the dynamics of the managerial planning and control processes for performance improvement. This hypothesis on the potentials for improving operations can be tested to some degree by the data presented in the Report itself. The Report analyzes the profitability of the companies acquired in relation to their industries. In 70 percent of the sample of acquired manufacturing firms, the profit rate of the acquired firm was below its industry profit rate. The Report also shows that the return on equity on the sample conglomerate firms it studied was 11.7 percent in 1969, not significantly different from the 11.5 percent for all manufacturing industries. Other evidence that the return on total assets or the return on equity of the conglomerate firms at the start of the conglomerate merger movement in the late 1950s was below the overall industry average provides some inferential evidence¹⁴ that the conglomerate firms have contributed to the efficiency of the combined operations.

■ The strongest position taken in the Report is that, as a consequence of conglomerate mergers, a serious loss of information has occurred.¹⁵ The requirement of product line reporting along four-digit SIC classifications by all firms is endorsed. One's initial reaction to such proposals is inclined to be positive since more information is usually better than less information. But more misleading information is not necessarily better than a smaller quantity of meaningful information. The multidivisional form of operations developed because of the economic gains from combining activities for which there are complementarities or carryovers—scale and efficiency effects.¹⁶ A division in a multiunit operation is held accountable for those activities over which its managers have control and therefore responsibility. Hence, managerial (internal) accounting procedures differ from financial reporting, because effective managerial control and allocation of resources do not require the development of overall profit measures for divisions comparable to those generated in an independent firm. To allocate various types of joint costs to individual divisions will be arbitrary at best and could be misinterpreted.¹⁷

Transactions are performed within an individual firm when they can be performed more efficiently than by market transactions.¹⁸ Thus, the theory of the firm (which includes team or organized effort¹⁹) also supports the rationale of a multiproduct, multiunit firm. The “*M*-form” has been further stimulated in recent years by the increase in the importance of the general management functions (planning, controlling, staffing) relative to the specific management functions (production, marketing).²⁰ Also, effective managerial planning and control activities are process-oriented rather than results-oriented.²¹ Hence, financial reports which are generally results-oriented are incomplete measures of performance at best. To

5. Numbers vs. information

¹⁴ See Weston and Mansinghka [16].

¹⁵ Cf. Mann [18] and Mueller [10].

¹⁶ See Williamson [18].

¹⁷ See Rappaport [11].

¹⁸ See Alchian [2] and Coase [7].

¹⁹ See Alchian [2], pp. 779, 793.

²⁰ See Weston [14].

²¹ See Weston [15].

make further arbitrary partitions would aggravate the distortions rather than provide more usable information.

Furthermore, much of the information required by public policy is revealed in the market place. This includes the prices, quality, etc., on individual products as well as data on inputs. In addition, information required by competitors or potential competitors is known in the trade. Thus, with respect to resource allocation decisions, the market research and other market intelligence operations of individual firms have developed very close estimates of "what is going on in the trade," as noted by Dr. Mann.²² The viability of the multidivisional firm depends upon achieving more efficient information flows and resource utilization than would be achieved by a number of independent firms performing the same operations. Thus, to require that an attempt be made to measure the "profitability" of interdependent divisions would be both unnecessary and misleading.

6. Conclusions

■ As the Report notes, diversification has long been an important attribute of U. S. firms. Diversification achieved by external acquisitions increased substantially during the 1960s. The major activity was centered in 60 to 70 firms with diverse backgrounds and uneven capabilities for carrying out their objectives.²³ The initial product base of the conglomerates strongly influenced their acquisition activity. In the sample of companies almost 60 percent of the acquisitions by number were product extension and 11 percent vertical. This confirms other evidence that diversification as such did not appear to be the primary objective of the conglomerate firms.²⁴ The opportunities available to the conglomerates were constrained.²⁵

An earlier major concern about conglomerates was their potential anticompetitive consequences. The Report concludes on this point as follows: "Finally, they do not appear to have become significant market forces in large sectors of our economy. They are located in a number of industries in which their individual market shares are relatively small and have shown no particular tendency to increase."²⁶ Nor are other apprehensions such as reciprocity, deep-pocket power, and forbearance supported by the evidence.

Since the conglomerates do not have market power, it is predictable that their profitability as a group will approximate the average for all industries. There is some evidence that they started below the average, so they have made some economic contribution. But their characteristics and constraints were highly diverse, so that dispersion in their efficiency and profitability achievements is substantial. This variability makes it difficult to establish as a generalization that potentials for synergy can be realized. Evidence suggests that diversification at a rapid pace by external acquisitions is more difficult to manage efficiently than a slower pace of internal diversification or a combination of the two.

The remaining issues of information flows relate more broadly than to conglomerate firms. Many operations performed in a single

²² In [8], p. 9.

²³ See Beman [3].

²⁴ See Weston, Smith, and Shrieves [17].

²⁵ See Weston [16].

²⁶ In [12], p. 199.

product firm could also be purchased from external markets. The proposals for "product line profit reporting" appear to be inconsistent with the underlying economic theory of the firm and its extension to multidivisional firms.

References

1. ADIZES, I. AND WESTON, J. F. "Comparative Models of Social Responsibility." *Academy of Management Journal*, Vol. 16, No. 1 (March 1973), pp. 112-128.
2. ALCHIAN, A. AND DEMSETZ, H. "Production, Information Costs, and Economic Organizations." *The American Economic Review*, Vol. 62, No. 5 (December 1972), pp. 777-795.
3. BEMAN, L. "What We Learned from the Great Merger Frenzy." *Fortune*, Vol. 87, No. 4 (April 1973), pp. 70-73ff.
4. BOCK, B. *Statistical Games and the '200 Largest' Industrials: 1954 and 1968*. New York: National Industrial Conference Board, Inc., 1970.
5. CELLER, E. *Investigation of Conglomerate Corporations Hearings* before the Antitrust Subcommittee of the Committee on the Judiciary, House of Representatives, 1969-1970. Parts 1-7.
6. CELLER COMMITTEE STAFF REPORT. *Investigation of Conglomerate Corporations*. June 1, 1971.
7. COASE, R. H. "The Nature of the Firm." *Economica*, Vol. 4, No. 16 (November 1937), pp. 386-405.
8. MANN, H. M. "Profit Reporting by Line of Business: A Question of How, Not When." Presentation to the Business and Industry Association of New Hampshire, November 16, 1972.
9. MARKHAM, J. W. "Market Structure and Decision-Making in the Large Diversified Firm," Ch. 14 in *The Impact of Large Firms on the U. S. Economy*, J. F. Weston and S. I. Ornstein, eds., Lexington, Mass.: D. C. Heath & Co., 1973.
10. MUELLER, W. F. "Corporate Disclosure: The Public's Right to Know." Paper presented to the Conference on Formulating Financial Reporting Standards, Northwestern University, October 18-19, 1971.
11. RAPPAPORT, A., FIRMIN, A., AND ZEFF, S. A. *Public Reporting by Conglomerates*. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1968.
12. U. S. FEDERAL TRADE COMMISSION. Staff Report: *Economic Report on Conglomerate Merger Performance*, November 1972.
13. ———. Staff Report: *Economic Report on Corporate Mergers*, August 28, 1969.
14. WESTON, J. F. "Discussion." *The American Economic Review*, Vol. 61, No. 3 (May 1971), pp. 125-127.
15. ———. "ROI Planning and Control." *Business Horizons*, Vol. 15, No. 4 (August 1972), pp. 35-42.
16. ——— AND MANSINGHKA, S. K. "Tests of the Efficiency Performance of Conglomerate Firms." *The Journal of Finance*, Vol. 26, No. 4 (September 1971), pp. 919-936.
17. WESTON, J. F., SMITH, K. V. AND SHRIEVES, R. E. "Conglomerate Performance Using the Capital Asset Pricing Model." *The Review of Economics and Statistics*, Vol. 54, No. 4 (November 1972), pp. 357-363.
18. WILLIAMSON, O. E. *Corporate Control and Business Behavior*. Englewood Cliffs, N. J.: Prentice Hall, Inc., 1970.
19. ———. "The Vertical Integration of Production: Market Failure Considerations." *The American Economic Review*, Vol. 61, No. 3 (May 1971), pp. 112-123.

Leonard Waverman's Natural Gas and National Policy: A Linear Programming Model of North American Natural Gas Flows

Reviewed by

David Kendrick

University of Texas at Austin

■ It was with a slight bit of surprise that this Texan realized today that the nearby natural gas fields here and in neighboring Louisiana are roughly 1000 miles closer to the major markets of eastern Canada than are Canada's own large gas fields in Alberta. However, this inescapable fact of geography cannot be avoided once one has watched Leonard Waverman use the lance of linear programming to attack Canadian natural gas policy makers.¹

What have the Canadian authorities done to attract Mr. Waverman's ire? They built expensive west to east pipelines across the rough terrain and long distances of Canada in order to keep cheap American gas out of eastern Canada instead of shipping Alberta's gas into California and letting Texas and Louisiana gas go into eastern Canada. He calculates the loss to Canada from this policy as 184 million dollars and concludes that this is too high a price to pay for nationalism.

How does he obtain this estimate? He solves the linear program

$$\text{Minimize } R = \sum_{i,j} (t_{ij} + p_i)x_{ij} \quad (1)$$

subject to:

$$\sum_i x_{ij} \geq D_j \quad (2)$$

$$\sum_j x_{ij} \leq S_i \quad (3)$$

$$x_{ij} \geq 0, \quad (4)$$

where

x_{ij} = Shipment of gas from field i to market area j ,

t_{ij} = Shipment cost,

p_i = The price of gas at field i ,

D_j = The demand for gas per day in market area j , and

S_i = The supply of gas per day available at field i .

He divides North America into nineteen producing areas including four in Canada, one in Mexico, and fourteen in the United States, and obtains estimates for the actual average daily production in these areas in 1966. Then he divides the United States and Canada

¹ L. Waverman, *Natural Gas and National Policy: A Linear Programming Model of North American Natural Gas Flows* (Toronto: Univ. of Toronto Press, 1972).

into nineteen market areas including nine in Canada and ten in the United States and calculates the average daily usage of gas in these areas in 1966. Field prices of natural gas in the United States and Canada for 1966 were obtained and used for p_i . Finally, he estimates a nonlinear transportation cost function, approximates it with fixed and variable cost elements, drops the fixed cost elements, and uses the variable cost elements with an adjustment made for the difficulty of the terrain to obtain the transportation cost elements t_{ij} . He then solves his problem to find the set of shipments which would have served the market demands at minimum cost per day in 1966.

In order to compute the cost of Canada's protectionist policy he then modifies the demand constraint from the form (2) to the form

$$\sum_i x_{ij} \geq D_j, \quad j \in A$$

$$x_{kj} \geq D_j, \quad j \in C, \quad (2a)$$

where

A = The set of American market areas,

C = The set of Canadian market areas, and

k = The production area index for Alberta.

The effect of this modification is to require that the eastern Canadian markets which are served largely by Louisiana and Texas fields in the first solution be served by Alberta via those Canadian pipelines in the protectionist solution. The cost for the solution to this second problem was about \$50.6 thousand dollars a day or \$18.4 million a year more than the cost for the free trade solution. The author then discounts this amount by 10 percent per year to obtain a present value of Canada's loss of \$184 million.

Mr. Waverman is the first to admit that his method of necessity employs a number of assumptions which most economists find unpalatable—namely fixed supplies, fixed demand, and linear transportation costs: all in a static model. The more appropriate model is obviously a multiperiod mixed integer programming model with demand functions, a capacitated pipe network and economies of scale in investment cost in the pipelines and (perhaps) in the exploration and development of new fields. However, the author performs a number of sensitivity tests on his models and concludes that none of them makes the Canadian pipelines economically competitive.

What he does not do is to tell us how much the \$184 million figure is modified by these sensitivity tests. If, as now appears to be the case, U. S. natural gas prices have been maintained at artificially low levels in recent years by the FPC, and if the loosening of these constraints results in a rise of U. S. relative to Canadian gas prices, this \$184 million figure could drop substantially.

What concerns me about this study the most is not the figure of \$184 million. The locational advantage of Louisiana and Texas both in crow flight distance and in difficulty of the terrain for pipelines convinces me that it is economically more efficient to service eastern Canadian markets from the United States rather than from Alberta. What bothers me is that the study looked backward rather than forward in time. The Canadian pipelines are sunk cost as is the rest of the existing pipeline network in North America. We should not

look back and ask whether or not those lines should have been built. Rather we should look forward by asking what shipments are optimal, given the pipelines which are already in place and the possibility of constructing new ones. This new problem should take seriously the fact that demand and supply are affected by prices of both natural gas and other energy sources, that existing pipelines have fixed capacities, and that new pipelines can be constructed. The resulting problem would probably be a mixed integer programming problem and would be difficult to solve. However, one can obtain bounds on such solutions which assure that the cost of the solution obtained is no more than such and such a percent greater than the global optimum. For my money I would rather have an approximate solution to the large problem than the exact solution to the smaller problem.

Jan Mossin's Theory of Financial Markets

Reviewed by

Mark E. Rubinstein

Assistant Professor of Finance
University of California, Berkeley

■ As one whose own work and education have been greatly influenced by Jan Mossin's ideas, I greet the opportunity to review Mossin's recent monograph¹ with some measure of humility. The admiring pupil must respectfully defer to the master. Those who have enjoyed the cogency and elegance of Mossin's published papers will not be disappointed. Where his monograph is redundant of his earlier work and where the related work of others is reviewed, the reader will find a fresh presentation with new insights; where new material is developed, Mossin preserves his characteristic laconic incisiveness.

The dominant theme—integration of portfolio theory, security valuation, financing-investment decisions of firms, and the optimality of security market allocation into a cohesive and logically satisfying model—reflects the modern development of financial theory under uncertainty, a dramatic burst of ideas fueled, to a large extent, by Mossin's own prior work. Mossin was among the first to perceive the implications of the mean-variance security valuation model and the objective of maximization of the value of corporate securities for corporate financial decisions. Like the classical construction of a play, Mossin moves from a brief Prologue (Chapter 1) into Act I, personal financial theory (Chapters 2 and 3), to Act II, security valuation (Chapters 4–6), and to Act III, corporate financial theory (Chapters 7 and 8), with the climax appropriately in Act II—statement of the mean-variance security valuation equation.

Chapter 2, "Risk Taking and Risk Aversion," contains one of the most readable and compact treatments of the relevant features of utility theory for financial analysis: the axiomatic basis of measurable utility, comments on Savage's simultaneous axiomatization of utility and subjective probability, the simplest version of the portfolio selection problem under uncertainty (one risk-free security with zero rate of return and one risky security with two states-of-the-world), implications of risk aversion, justification and insurance analysis of absolute and proportional risk aversion,² comments on indirect utility functions, and the significance of the distinction between temporal and timeless prospects in the "two-period" consumption allocation problem.³ This would also have been an appropriate place for a development of temporal measures of risk aversion,⁴

¹ J. Mossin, *Theory of Financial Markets* (Englewood Cliffs, N. J.: Prentice-Hall, 1973).

² This is less complex than Pratt [16].

³ Mossin modestly neglects to cite his own paper [14].

⁴ As in Dreze and Modigliani [6].

multiperiod portfolio models which maximize expected utility of terminal wealth,⁶ and extension to withdrawals for intermediate consumption.⁶ However, presumably to preserve the character of the monograph in cohesiveness of subject matter, these extensions are omitted in favor of peppered algebraic (quadratic and logarithmic utility) and numerical illustrations, reinforcing the rigorously covered material.

Chapter 3 generalizes the simple portfolio selection model to arbitrary probability distributions, a nonzero risk-free rate, and an arbitrary number of securities. Similarly to Arrow,⁷ Mossin proves the portfolio implications of absolute and proportional risk aversion for choice of risky securities. Mossin notes that $u(Y)$ and $u(R)$, where Y is end-of-period wealth and R is the portfolio rate of return, are equivalent as utility functions only with constant proportional risk aversion. As Markowitz⁸ worked with $u(R)$, Mossin⁹ has uncovered a minor flaw in Markowitz' analysis. While not mentioned in the monograph, it is esthetically pleasing to note here that $u(Y)$ and $u(Y - W)$, where W is initial wealth so that $Y - W$ is income, are equivalent as utility functions only with constant absolute risk aversion. As a lone and parting shot at the dynamic portfolio problem, Mossin mentions briefly the surprising constellation of properties which bless the objective of maximization, after many periods, of the expected logarithm of terminal wealth. In view of these properties, as well as others, I was disappointed at the omission of a more thorough analysis of a problem that has already enlisted the formidable combination of Hakansson, Markowitz, Merton, Samuelson, and Thorp in intellectual combat. With this cautious dalliance, Mossin returns to the more mundane features of his nondynamic portfolio problem. He shows that realistic restrictions on the use of proceeds from short sales imply a *range* of prices over which zero holding of the risky security is optimal. He contrasts his single-stage procedure to the Markowitz two-stage (efficient set) approach, which he aptly characterizes as "computing first and thinking afterward." He then leaves the portfolio efficiency approach behind him. However, in view of the recent generalization of it with concepts of stochastic dominance,¹⁰ such short shrift is regrettable. Under quadratic utility, a favorite simplification, Mossin shows quite simply that it is the portfolio separation property which effectively permits the single-stage analysis. At this point, to clarify the results of Chapters 5 and 6, Mossin should have derived the entire class of utility functions for which portfolio separation holds¹¹; unfortunately, he omits this generalization. In view of Cass and Stiglitz' work,¹² his discussion of the extent to which the results of the two-security case carry over to arbitrary numbers of securities is similarly incomplete. To conclude the chapter, after a brief discussion of the Markowitz-Sharpe "single-factor market model," he fires a well-aimed salvo at the

⁶ Developed initially by Mossin, himself [13].

⁶ As in Hakansson [8].

⁷ In [2].

⁸ In [10].

⁹ In [13].

¹⁰ In Hanoeh and Levy [9].

¹¹ See Cass and Stiglitz [4].

¹² In [3].

“stable-Paretian hypothesis:” just because the probability distributions of rates of return on individual securities appear to have (for purposes of approximation) infinite variance, does not mean that this holds for portfolios. Indeed, Mossin’s own empirical test shows portfolio kurtosis of rate of return falls rapidly to three (the kurtosis of the normal distribution) as the number of securities in randomly selected portfolios is increased. While it is hardly appropriate here to survey this contentious area, perhaps a more fundamental reason for rejecting the “stable-Paretian hypothesis” is the insufficient temporal stationarity of the probability distribution of security rates of return.¹³

Chapter 4 follows fairly closely the derivations of the mean-variance security valuation model given by Mossin.¹⁴ Mossin proves that with quadratic utility the “market price of risk” equals the inverse of the sum of the expected inverses of individual absolute risk aversions. Left unanswered is whether or not this result carries over more generally to the mean-variance security valuation model without quadratic utility, but with normal probability distributions. It would therefore seem appropriate to state here, that except for exponential utility, it does not.¹⁵ Impatient to turn to corporate financial theory, Mossin examines the most important implications of the model for capital structure and mergers. Appropriately stressing the assumption of fixed operating income independent of capital structure, he intuits the central Modigliani-Miller¹⁶ result as well as non-synergistic merger irrelevancy. While he is seldom given credit, Mossin¹⁷ was the first to rigorously prove this latter irrelevancy proposition. He closes the chapter with some practical considerations: why individuals do not seem to diversify so much as the model implies—implications for mutual fund management and random portfolio selection.

Chapter 5, “Extensions of the Valuation Model,” is in my view the “least best” of the monograph. Attempting to free the capital structure irrelevancy proposition from the restrictive mean-variance context, Mossin purports to prove that it holds generally even in the presence of default risk on bonds. To this end, he first proves his “additivity proposition:”

In equilibrium, the market value of the sum of the returns on a set of securities is equal to the sum of the market values of the returns on those securities.

As applied to a single economy, this proposition is unobjectionable; however, as applied to two otherwise similar economies (the first in which only a portfolio, and not its constituent implicit securities, can be purchased and the second in which each constituent security can be purchased separately), it is generally false. In general, the ability to purchase a greater variety of securities, even though the probability distribution of future social wealth is unaffected, enlarges relevant exchange opportunities and affects equilibrium prices. Consequently, if a firm levers itself so as to have a positive probability of default, it will create a new and desired pattern of returns across

¹³ See Rosenberg [17].

¹⁴ In [12, 15].

¹⁵ See Rubinstein [18].

¹⁶ In [11].

¹⁷ In [12].

states, thereby enlarging exchange opportunities and affecting its value. As Stiglitz¹⁸ has shown, this conclusion can be expected to hold except for the special cases of (1) a complete (Arrow-Debreu) securities market or (2) the existence of portfolio separation where all individuals purchase the *same* risky mutual fund. With (1) all possible patterns of security returns are already available, and with (2) any new patterns (with future social wealth fixed) are irrelevant to decisions since all individuals simply purchase the next portfolio of risky securities. In these special cases the “additivity proposition” holds across economies, and consequently capital structure will be irrelevant even with default risk. Returning to the mean-variance model, by assuming a convenient two-point probability distribution of bond returns, Mossin derives the equilibrium contracted, as opposed to expected, rate of return on bonds. Then follows a more traditional sequence of assumption relaxation: risky borrowing and lending and heterogeneity of beliefs. The most regrettable omission, particularly in view of the recent work of Hahn, Kurz, Shubik¹⁹ and Starr, is a treatment of transactions costs, the role of money, and game-theoretic aspects of market equilibrium. A general pedagogic criticism is that his presentation in this chapter would have been clearer and more compact if he had developed his most general mean-variance model first and then illustrated the role of assumptions by tightening instead of relaxing them.

Chapter 6, “Efficiency of Security Markets,” an investigation of the Pareto-optimality of risk sharing among individuals, was for me the highlight of the monograph. For background, Mossin first proves the well-known necessary and sufficient condition for Pareto-optimality under certainty and then the Pareto-optimality of a competitive allocation of commodities. This, of course, is the fundamental justification of a competitive economy. The question remains whether or not this crucial property carries over to a competitive securities market under uncertainty with homogeneous beliefs. Mossin concludes that Pareto-optimality does not carry over except for two special cases:

- (1) a complete securities market, or
- (2) all preferences satisfy $\frac{-u_i'(Y_i)}{u_i''(Y_i)} = \alpha_i + \beta Y_i$,

where i represents individuals. The first case is familiar from Arrow and the second from Wilson.²⁰ Solving the differential equation, case (2) is restrictive but includes most popular utility functions, such as the quadratic, exponential, logarithmic, power, etc. Mossin, however, fails to exploit the intuitive significance of case (2). Not coincidentally—in the absence of restrictions on probability distributions—(2) is also the necessary and sufficient condition for all individuals to choose the *same* two mutual funds, one risk-free and the other the *market* portfolio of risky securities.²¹ They exhibit this conforming behavior irrespective of the variety of risky securities, even if exchange opportunities were enlarged to the limit of complete

¹⁸ In [21, 22].

¹⁹ In [20].

²⁰ In [1] and [23], respectively. Mossin’s assumption of belief homogeneity is unnecessary for case (1).

²¹ See Cass and Stiglitz [4].

markets. Constraining exchange, as Mossin does, to the linear sharing rule $Y_i = a_i + b_i W$, where W is future social wealth, does not impinge on *relevant* exchange opportunities, since even with the more general competitive sharing rule $Y_i = a_i + \sum_j b_{ij} X_j$, where X_j is the future value of security j , and with enough securities for complete markets, each individual *chooses* $b_{ij} = b_i$ for all securities. Thereby, the general competitive sharing rule collapses to $Y_i = a_i + b_i W$, where $W = \sum_j X_j$. Consequently, since (1) is Pareto-optimal, then with portfolio separation (even without complete markets), (2) must also be Pareto-optimal. While Mossin misses this convenient interpretation, his proof of Pareto-optimality is more elegant than the corresponding Cass and Stiglitz²² proof of portfolio separation. Mossin also neglects a third and empirically significant case in which the Pareto-optimality of risk sharing holds up in a competitive securities market: all securities have normally distributed rates of return. The generalized separability conditions of Cass and Stiglitz²³—with some restrictions to insure all individuals select the *same* mutual funds—may constitute a fourth instance. However, generally risk sharing will not be Pareto-optimal—the major block to optimality being a combination of incomplete markets and heterogeneous beliefs.

In Chapter 7, Mossin examines corporate financial decisions in greater depth, in particular, capital budgeting and the Pareto-optimality of value maximization for the allocation of investment among firms. He criticizes the Modigliani-Miller analysis for coping rigorously only with nondiversifying investments, a restriction remedied in the mean-variance model. He notes that while all firms within the same risk class must have the same rates of return, if one firm invests in additional assets within the risk class (nondiversifying), the rates of return of *both* firms may change. Moreover, a similar argument follows for capital structure: just because two firms in the same risk class have the same value does not imply that this common value will be unaffected by their capital structures. Implicitly, Modigliani and Miller ignored this second-order effect by a partial equilibrium analysis which assumes the decisions of a firm do not affect the decisions or capitalized values of other firms. In contrast to the Modigliani-Miller analysis, Mossin argues that for diversifying investments, the cost of capital concept loses its meaning since the cut-off expected rate of return will be *different for each project*. However, as I have noted elsewhere,²⁴ the marginal cost of capital of a project may be interpreted as the expected rate of return demanded by individuals on securities with the same risk as the project. However, a complicating factor to which Mossin draws attention, is that in general the cut-off expected rate of return (for value-maximizing firms) will be *different for each firm*. That is, a firm will measure the riskiness of a project by the sum of (1) the project's covariances with all firms including itself, (2) the project's own-variance, and (3) the project's covariance with the firm. Clearly, (3) will be different for different firms even for the same project. However, with a reasonably diverse and large economy, both (2) and (3) should be very

²² *Ibid.*

²³ *Ibid.*

²⁴ In [19].

small relative to (1) for most projects. Consequently, to a good second-order approximation, the cut-off expected rate of return for a project will be the same for all firms. Mossin neglects to specify the likely relative sizes of (1), (2), and (3), and therefore is led to over-emphasize the differences in project acceptance criteria among firms. To the extent that (3) looms relatively large, the allocation of investment among value-maximizing firms cannot be Pareto-optimal, since it follows from portfolio separation that individuals are indifferent as to which firm accepts a nonsynergistic project. Another important divergence from Pareto-optimality which Mossin mentions is the failure of a value-maximizing firm to fully internalize the riskiness of its investments; other firms with rates of return positively correlated with the rate of return of an accepted project will experience a decline in value, since from a social point of view their output becomes more risky. This in turn may cause them to revise their own capital budgeting decisions.

In the final chapter, Mossin moves from his comparative statics analysis of corporate financial decisions to Diamond's model²⁵ in which the demand and supply of securities are simultaneously determined in equilibrium. He also considers the Pareto-optimal implications of changes in the "market price of risk" which he has hitherto held fixed in his comparative statics analysis. Recently, Fama²⁶ has published a more comprehensive and carefully reasoned treatment of the Pareto-optimality of the allocation of investment in a competitive economy.

If my comments have appeared excessively critical, it is only because of the high standards against which Mossin's work deserves to be assessed. The development of financial theory (and supportive empirical work) has moved with such rapidity in recent years that it is hardly surprising that every chapter of the 1973 monograph is no longer completely up to date.²⁷ To preserve its attractive length and merely to bring the work to an end, many fast breaking extensions of the theory had to be given short due or omitted entirely. Perhaps most impressive, Mossin covers most of the foregoing and more²⁸ on both a mathematically rigorous and verbally intuitive level in only 164 pages! In the light of current knowledge, Mossin's only significant flaw is his failure to realize the full implications of portfolio separation. From a more recent unpublished paper of Mossin, evidently he himself now perceives this same defect.

A final pedagogical note: given the level of the mathematics (a smattering of calculus, linear algebra, and probability theory) and the depth of its verbal abstractions, Mossin's own assessment—a concise "core" text around which to organize an advanced graduate course—seems justified. I certainly found the monograph instructive, with its natural place in my library along side the works of Markowitz and Fama-Miller.

²⁵ In [5].

²⁶ In [7].

²⁷ See, for example, Cass and Stiglitz [3, 4], Fama [7], and Stiglitz [21, 22].

²⁸ My review of Chapters 7 and 8 was very spotty.

References

1. ARROW, K. J. "The Role of Securities in the Optimal Allocation of Risk-Bearing," *Review of Economic Studies*, Vol. 31, No. 3 (April 1964), pp. 91-96.

2. ———. "The Theory of Risk Aversion," in *Essays in the Theory of Risk-Bearing*, Chicago: Markham, 1971.
3. CASS, D. AND STIGLITZ, J. E. "Risk Aversion and Wealth Effects on Portfolios with Many Assets," *Review of Economic Studies*, Vol. 39, No. 3 (July 1972), pp. 331-354.
4. ———. "The Structure of Investor Preferences and Asset Returns, and Separability in Portfolio Allocation: A Contribution to the Pure Theory of Mutual Funds," *Journal of Economic Theory*, Vol. 2, No. 2 (June 1970), pp. 122-160.
5. DIAMOND, P. A. "The Role of the Stock Market in a General Equilibrium Model with Technological Uncertainty," *The American Economic Review*, Vol. 57, No. 4 (September 1967), pp. 759-776.
6. DREZE, J. H. AND MODIGLIANI, F. "Consumption Decisions under Uncertainty," *Journal of Economic Theory*, Vol. 5, No. 3 (December 1972), pp. 308-335.
7. FAMA, E. F. "Perfect Competition and Optimal Production Decisions under Uncertainty," *The Bell Journal of Economics and Management Science*, Vol. 3, No. 2 (Autumn 1972), pp. 509-530.
8. HAKANSSON, N. H. "Optimal Investment and Consumption Strategies under Risk for a Class of Utility Functions," *Econometrica*, Vol. 3, No. 5 (September 1970), pp. 587-607.
9. HANOC, G. AND LEVY, H. "The Efficiency Analysis of Choices Involving Risk," *Review of Economic Studies*, Vol. 36, No. 3 (July 1969), pp. 335-346.
10. MARKOWITZ, H. M. *Portfolio Selection: Efficient Diversification of Investments*. New Haven: Yale Univ. Press, 1970.
11. MODIGLIANI, F. AND MILLER, M. H. "The Cost of Capital, Corporation Finance, and the Theory of Investment," *The American Economic Review*, Vol. 48, No. 3 (June 1958), pp. 261-297.
12. MOSSIN, J. "Equilibrium in a Capital Asset Market," *Econometrica*, Vol. 34, No. 4 (October 1966), pp. 768-783.
13. ———. "Optimal Multiperiod Portfolio Policies," *Journal of Business*, Vol. 4, No. 2 (April 1968), pp. 215-229.
14. ———. "A Note on Uncertainty and Preferences in a Temporal Context," *The American Economic Review*, Vol. 59, No. 1 (March 1969), pp. 172-174.
15. ———. "Security Pricing and Investment Criteria in Competitive Markets," *The American Economic Review*, Vol. 59, No. 5 (December 1969), pp. 749-756.
16. PRATT, J. "Risk Aversion in the Small and in the Large," *Econometrica*, Vol. 32, No. 1-2 (January-April 1964), pp. 122-136.
17. ROSENBERG, B. "The Behavior of Random Variables with Nonstationary Variance and the Distribution of Security Prices." Unpublished manuscript, University of California, Berkeley, 1972.
18. RUBINSTEIN, M. E. "A Comparative Statics Analysis of Risk Premiums." Forthcoming in *Journal of Business*.
19. ———. "A Mean-Variance Synthesis of Corporate Financial Theory," *Journal of Finance*, Vol. 28, No. 1 (March 1973), pp. 167-181.
20. SHUBIK, M. "A Theory of Money and Financial Institutions: Parts I-IX." Cowles Foundation Discussion Papers #320, #322, #324, #330, #331, #334, #336, #337, #338, 1971-1972.
21. STIGLITZ, J. E. "On Some Aspects of the Pure Theory of Corporate Finance, Bankruptcies and Take-Over," *The Bell Journal of Economics and Management Science*, Vol. 3, No. 2 (Autumn 1972), pp. 458-482.
22. ———. "A Re-Examination of the Modigliani-Miller Theorem," *The American Economic Review*, Vol. 59, No. 5 (December 1969), pp. 784-793.
23. WILSON, R. "The Theory of Syndicates," *Econometrica*, Vol. 36, No. 1 (January 1968), pp. 119-132.

OMISSIO

■ In the article, "Depreciation in a Simple Regulatory Model," by Bruce L. Jaffee, *The Bell Journal of Economics and Management Science*, Vol. 4, No. 1 (Spring 1973), the following figure should appear in the right-hand margin on page 341.

FIGURE 1

TWO POSSIBLE SHAPES FOR THE
OPTIMAL PATH OF DEPRECIATION

